Consultation Feedback

on the Draft AI Ethics Guidelines published by the High-Level Expert Group on Artificial Intelligence on 18 December 2018

First name	*Last name*	*Organisation*	*Introduction: Rationale and Foresight of the Guidelines*	*Chapter I: Respecting Fundamental Rights, Principles and Values - Ethical Purpose*	*Chapter II: Realising Trustworthy AI*	*Chapter III: Assessing Trustworthy AI*	*General Comments*
Miles	Brundage	OpenAI		There was a lot of useful information in this section, but I felt it perhaps leaned a bit heavily on the AI4People's synthesis of ethical frameworks - without readers doing some combination of reading the footnotes and reading the original paper, it's not clear whence the apparent biomedical ethics analogy/transfer, what key ethical constraints/risks if any might be obscured by this framework, etc. Section 5 on "Critical concerns raised by AI", regarding the controversy over this section, it wasn't clear to me what the source of controversy is. If it is whether these risks are worth considering whatsoever, then it seems the answer is obviously "yes" - none can be definitively ruled out. If the question is whether they are the most important risks, then in my opinion, the answer is unclear - if I were to make such a list, other risks would be included (e.g. economic impacts; malicious uses of AI in the cyber realm; etc.). Furthermore, I'd suggest changes to the section on potential longer term concerns (which I think should probably be kept, but with tweaks). In particular, while flagged as "controversial," notably some of the points made here have been agreed upon by a fairly diverse set of actors who signed the Asilomar Principles (principles 6, 10, and especially 19 seem relevant). My guess is that dropping a few of the buzzwords/phrases such as unsupervised recursively improving AGI and artificial consciousness might make this section read more convincingly/uncontroversially. You might also consider citing Grace et al.' 2017's survey of AI expert opinion, noting disagreement among experts regarding the long term development of AI. Finally, the authors might generally wish to be clear on uses of terms like safety, robustness, value alignment, etc. See e.g. this suggested lexicon for consideration https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1	See above regarding precision of language on robustness, safety, etc. - the same also applies to reliability and reproducibility, and "resilience to attack." Regarding the safety section, see above regarding the value of using precise language, and consider citing recent work surveying these areas or touching on topics mentioned in this report, e.g. Concrete Problems in AI Safety as an overview and AI Safety Gridworlds as an example of the standardized evaluation mentioned. The mention of privacy/security by design seemed a bit strange to me - while privacy by design is a commonly mentioned concept, I have not yet heard security by design mentioned in the context of AI. If we are to go in such a direction, why not robustness-by-design, or safety-by-design? Are these different? What does security mean here? Also, the idea of both a fail-safe and a reboot mechanism are both pretty specific proposals, which is good to see, but I'm not sure the evidence/theory base is sufficient to justify adopting these at large scale. Such a requirement would seem to be hard to specify/implement and potentially unhelpful/costly in some cases, so I'm not sure I'd support this being specified yet. Re: the testing and validation section, note that there is a substantial literature on these topics which might be referenced.	In the section with bullets, the robustness section seems very short, and again, not totally distinct from safety and other areas (reliability & reproducibility; governing AI autonomy) - you might consider rearranging these to be more even in length, and avoid real or perceived redundancy.	

Anonymous	Anonymous	Anonymous		For section 5.4, I think the matter of possible AI arms races is significant enough that it deserves a point of its own. It is important that AI projects can set out safety rules, and can then inspect that the other project is following those rules, without being able to steal each others' work. If parties are in an arms race where they can violate safety norms without sanctioning, then safety might be flouted in order to make slightly faster progress, leading potentially to widespread harm (in violation of the principle of non-maleficence). More on arms races here: https://www.fhi.ox.ac.uk/wp-content/uploads/Racing-to-the-precipice-a-model-of-artificial-intelligence-development.pdf In section 5.5, the use of the phrase "unsupervised" in "unsupervised recursively self-improving AGI" seems incorrect to me. "Unsupervised" in an AI setting is a technical term that refers to unsupervised learning. A recursively self-improving AI in-principle could use supervised (using a dataset of AI programs and their outputs) or unsupervised learning, and could pose similar threats in either case, so the word should be removed.		For 4. Governing AI Autonomy section, I think it would be useful to also state the technical names for these desired properties: "to allow human control, if needed, in each state" is called "having a human in the loop". The "stop button" is known in the technical literature as "interruptibility" or "corrigibility". For "9. Safety", it would be useful to make the point that the amount of effort taken to ensure safety should scale with the magnitude of the safety threat. If serious longer-term threats do arise (although we may disagree on the likelihood of this), then we would agree that the safety measures taken should be commensurate to that risk.	
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Anonymous	Anonymous	Anonymous					(A) I saw that these guidelines are voluntary, and they're also pretty vague. A.I can be very dangerous and large corporations can use it for their own good, leaving behind the common people. I think that next to these loose guidelines, there should be strict rules written into EU law to ensure the safety of our citizens. (B) I've seen that the document does not account for the possibility of superintelligence in the future. This is very important in A.I guidelines, as superintelligences are the most dangerous. For example, to avoid big catastrophes with super intelligence, it is important that AI systems never have access to the internet, but merely to a few local databases that get updated when needed. This is extremely important to ensure safety in our A.I.
Ryan	Vannin			Amend 5.2 (p. 11): From "[...] AI developers and deployers should therefore ensure that humans are made aware of – or able to request and validate the fact that – they interact with an AI identity." to "[...] AI developers and deployers should therefore ensure that humans are made aware of they interact with an AI identity by issuing clear and transparent disclaimers." (Forcing AI developers and deployers to clearly and transparently inform users that there's an interaction with an AI entity complies better with the aims and goals of current regulation, such as GDPR; Moreover, to a user must always be given the option to opt-out of a human-machine interaction in	Addition in 2. (p. 22): " - Transparency and openness" Organisations deploying AI systems shall inform their users when and where decisions are taken by AI, respectively by humans. Current advances in technology can mislead a user by letting him think that he's currently interacting with a human (see Google's bot called Duplex with a humanlike voice calling staff at a restaurant and hair salon to make reservations). This shall be disclosed at the beginning of any human-machine interaction and be always given the possibility to opt-out in favour of a human-human interaction.	

			favour of a human-human one).	
Anonymous	Anonymous	Anonymous	See section on Explanation (XAI) - PDF page 21. I see a potential danger with XAI in normative decision making that should perhaps be called out. For illustration, consider an extreme case in which a medical AI decides that it is best to execute one healthy patient and harvest their organs in order to save five other patients requiring transplants. Assume that the AI explains its reasoning in terms of (say) maximizing wellbeing. Here are three ways that a human operator might respond to this: 1) Simply disagree with the AI in this case and override it. 2) Assume that the AI is defective and send it for retraining. 3) Be convinced that the AI is correct and that one patient should indeed be executed to save five others. How do we avoid (3)? It seems easy to dismiss in an extreme example such as this one, but what if it were a more realistic and nuanced scenario? The situation is analogous to the perfectly common case of a human accepting moral testimony from a convincing, yet morally defective individual (imagine a case where a charismatic dictator turns public sentiment against some minority group). So we are left with a dilemma: if the explanation provided by XAI cannot convince us to change our minds about ethical decisions, then the explanation is redundant. But if it can, then it is potentially dangerous. Suppose again that an XAI presents an operator with an explained decision. The space of possibilities is: a) The decision is right for the reasons explained. Example: “The patient must not be executed based on a principle of non-maleficence.” b) The decision is right but not for the reasons explained. Example: “The patient must not be executed as it would result in a lawsuit.” c) The decision is wrong for the reasons explained. Example: “The patient must be executed based on a principle of non-maleficence.” d) The decision is wrong but not for the reasons explained. Example: “The patient must be executed because today is Monday.” In (c) and (d) the operator will (presumably) consider the AI straightforwardly defective. In (b), the AI has made the right decision for the wrong reasons, so _should_ be considered defective. However, there is a chance that the operator may not consider it defective, because after all it has the right result, and/or humans generally have poor moral reasoning capabilities. In (a), the operator will (presumably)	

					consider the AI non-defective. But this could be due to a convincing explanation provided by the XAI even though the decision is in fact morally wrong (since we must assume that the AI, if it is to be useful and trusted, is capable of making decisions at least as well as we can). This is - I think - a difficult and important problem for XAI. It has clear parallels with the problem of moral testimony in the philosophical field of metaethics, so I suspect there is scope for overlapping research here.		
Anders	Arpteg	Peltarion	Nice overview of the motivation and need for ethical guidelines. One sentence that caught my eye was "it should be noted that no legal vacuum currently exists, as Europe already has regulation in place that applies to AI". To me, this sounds like there is no need to update / review / adapt the current regulation to an AI-First future. If this is the intent, that would be questionable as our regulation is in clear need to be updated continuously to make sure we maximize the benefits and minimize the risks of AI.	Certainly agree with that a human-centered approach is the best way forward short term. Would be interesting to also consider from a more long-term perspective. There are people (Elon Musk and others) that believes it could be worth considering a future where the machine is generally more intelligent that humans (perhaps 30+ years forward), and what the best way to maximize the benefits of AI would be in such a world.	Great overview of technical and non-technical concerns in realising trustworthy AI. From an objective point of view, it would also be interesting to consider an abuse / overuse of ethical concerns. Could it have negative effects in terms of maximizing the benefits of AI, e.g. companies starting to limit storage of historical data due to misunderstanding of current regulation?	In consideration of having as many companies adopting these guidelines as possible, having clear and concrete recommendations and "assessment lists" is of high importance IMHO. To make the list as concise as concrete as possible, perhaps some items can be merged as the are somewhat overlapping. For example, is it necessary to have both "design for all" and "non-discrimination" as separate items?	Great set of guidelines, good work so far. In general, to make sure that we truly maximize the benefits of AI and minimize the risks, it is important that we also incorporate potential downsides with over-regulation, legal uncertainty consequences, or misunderstandings of how to make the best use of this set of guidelines. For example, it can be worth talking about all the technical advances to improve privacy (eg. differential privacy, homomorphic encryption, split NNs), providing clear incentives for companies to self-regulate, and that an improperly defined set of regulation could have negative effects.Another small note, the consultation form works in Firefox but does not seems to work in the Google Chrome web browser.
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
john	hunt	Internet					My hope is that you will imbue AI and robots with the ethics of peace. Peace requires NO INITIATION OF FORCE against a human by another human. As AI learns by watching humans, it needs to watch humans who hold to the ethics of NO INITIATION OF FORCE. If they, instead, observe socialists and fascists as their models, they will learn from people who believe it is okay to initiate force against other humans. If AI learns that this is okay, then it will quite readily initiate force against humans. The ethics of peace are not the ethics of socialism and statism. The ethics of peace are simple: 1) Don't initiate force or fraud against another person. 2) Do what you agree to do.

							My recommendation is simple. Teach AI what you can find at www.ethicssolutions.net. Teach them this. Let them watch people who ascribe to this ethics. Best John Hunt, MD
Dr Karl	Gosejacob	GOSEJACOB & Bundesverband der deutschen Industrie (BDI)	A certain skepticism towards AI is more than justified, thus should be parts of the ethics. In a sense, AI is automating experience rather than understanding or proof – i.e. you never know why AI comes up with a certain result. AI, since the term was coined in the 1950s, is quite a dubious concept, compared especially to mathematics. As of new, there are too many AI gold diggers around.	Again, more skepticism is needed. IBM's Watson, is not doing that well in medicine, is it/he/she? – This chapter seems to be more on the Asimov science fiction side of view, which can be entertaining. –	Core should be traceability, repeatability and peer review recognition, as in maths, medicine, pharmaceutical research, or physics. AI should always have to pass experimental tests, i.e. should be facts-based. The claim 'being AI' is too much of just advertising.	See Chapter II.	Don't just rely too much on self-declared AI experts, I hesitate to call them evangelists.
				Section B I , chapter 4: Ethical Principles in the Context of AI and Correlating Values Page 8 ff (also referring to page 11, number 5: Critical concerns raised by AI): Consider a 6th principle: "The principle of liability: Act Responsibly Be it researchers, designers, producers, vendors or users of AI systems, they shall be aware of their responsibility for what they do and ready to accept their liability for the implications."				
Norbert	JASTROCH		Executive Summary Page i, 3rd paragraph: Clarify 'Human-centric approach to AI' by inserting … to increase human well-being. "Putting the human being in the center calls for the requirement to protect personal integrity, respect individual liberty, and convey generic diversity." Trustworthy AI will be our north-star ... Page iv, Human-centric AI: … to increase human well-being. "Putting the human being in the center calls for the requirement of protecting personal integrity, respecting individual liberty, and conveying generic diversity."	Page 12, chapter 5.3: Normative & Mass Citizen Scoring without consent in deviation of Fundamental Rights: Include the value of diversity of individuals, and add the danger of private organisations installing mass scoring: We value the freedom and autonomy of all citizens, "as well as the diversity of individuals which is source of creativity, innovation and societal development". Normative citizen scoring (e.g., general assessment of 'moral personality' or 'ethical integrity') in all aspects and on a large scale by public authorities "or private organisations" endangers ...“ Page 12/13, chapter 5.5: Potential longer-term concerns. Clarify/exemplify by inserting ... to the very distant future). "However, one research direction is already being pursued that aims at the develpoment of bio-technical interfaces between brain and external devices. Examples are bionic applications for the restitution of brain control over (prothetic) parts of the body, or research into possibillties to bypass the locked-in syndrom. While these are ethically well acceptable, they might be extended to cyborg-type experiments (the linking of the human brain to a computer) that would raise all kinds of ethicallly questionable implications like the dissolution of personal integrity, autonomy and responsibility." A	Section B II , page 22: Codes of Conduct: Consider the formulation of an "EU code of conduct for research into AI."	Section B III: Assessing trustworthy AI: Further to the assessment list, consider "the development and maintenance of an EU code of conduct for research into, development and application of AI, combined with the introduction of a respective label 'EU trusted AI'."	Congratulations for an extraordinary piece of work.

				risk-assessment approach therefore … .				
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	
Nomi	Byström	Aalto University	See General Comments	See General Comments	See General Comments			I recommend that the Ethics Guidelines, in addition to the Charter of Fundamental Rights of the European Union, also include reference to the European Convention on Human Rights (ECHR/Convention). Accordingly, it would be significant for Trustworthy AI made in Europe to refer to both fundamental and human rights, for the following reasons: First, adding reference to both the fundamental rights of the Charter and the human rights guaranteed by the ECHR is justified by the fact that the European Union's fundamental values include respect for human rights. Inclusion is also strengthened by the statement in the draft ethics guidelines (p. 1): 'AI is thus not an end in itself, but rather a means to increase individual and societal well-being.'Moreover, the Charter and the ECHR have a strong link with a recognised correspondence that goes way beyond Article 6(2) of the Treaty of Lisbon. The connection is evident in the explanations prepared and updated under the authority of the Praesidium of the Convention that drafted the Charter. To give just one example from the Official Journal of the European Union (2007) C303/17:'Explanation on Article 10 — Freedom of thought, conscience and religionThe right guaranteed in paragraph 1 corresponds to the right guaranteed in Article 9 of the ECHR and, in accordance with Article 52(3) of the Charter, has the same meaning and scope. Limitations must therefore respect Article 9(2) of the Convention, which reads as follows: "Freedom to manifest one's religion or beliefs shall be subject only to such limitations as are prescribed by law and are necessary in a democratic society in the interests of public safety, for the protection of public order, health or morals, or for the protection of the rights and freedoms of others."'*Furthermore, by referring to the human rights guaranteed by the European Convention on Human Rights, the position of Trustworthy AI made in Europe is

							strengthened. This is due to the fact that there are limits to the applicability of the Charter that the ECHR is free from. Regarding non-discrimination: (See, for example, Chapter I (3): Fundamental Rights of Human Beings and Chapter II: Realising Trustworthy AI) Of particular relevance for the requirements of Trustworthy AI, its number five: non-discrimination is Protocol No. 12 to the ECHR. According to its Article 1: 'General prohibition of discrimination1.The enjoyment of any right set forth by law shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.2. No one shall be discriminated against by any public authority on any ground such as those mentioned in paragraph 1.'** *https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.C_.2007.303.01.0017.01.ENG&toc=OJ:C:2007:303:TOC ** https://www.echr.coe.int/Documents/Convention_ENG.pdf
Paris	Mens	MensArtis	pp.i. ref.: "Trustworthy AI has two components: (1) it should respect fundamental rights, applicable regulation and core principles and values, ensuring an "ethical purpose"" => "Respect" refers to AI as an agent, however A.I. is a technology or an industry and therefor it should not only respect such, it should implement such: Trustworthy AI has two components: (1) it should >implement< fundamental rights, applicable regulation and core principles and values, ensuring an "ethical purpose" pp.ii Consider technical and non-technical methods to ensure the implementation of those requirements into the AI system. seems correct but pp iii Adopt an assessment list for Trustworthy AI when developing, deploying or using AI, Seems to imply that the assessment is post hoc at every step, indeed the word "designing" should be added wherever the above list is used. Indeed the terms "design stage" and "architecture stage" are missing from the chapter. "design is mentioned on page 2, but not in "All relevant stakeholders that develop, deploy or use AI" (even though software development is practically primarily incremental design). Design/Architecture are mentioned in chapter 2. As for rationale, in general; we as Western society, by now had multiple waves of technological work replacement and hence have the opportunity to learn from them, in order to provide guidelines a priori and even technologies to feed the design of systems. We have to learn from the past.	3. Fundamental Rights of Human Beings (pp.7) The main problem that can be expected from the introduction of the most impactful A.I. technology (self driving cars) is that an army of truck drivers lose their jobs. But this change is not unprecedented: retail employees with human interaction skills are in large numbers being replaced by warehouse employees that fill order carts or in turn by robots that do that. AI has similarly made the optimisation of labor costs possibly result in uberisation: by cutting out the positions that feed back employee concerns into the organisation (replacing those with "Help" or "Q&A" sections of websites or by waivers in "informed consent" buttons). Paragraph 3 completely misses these problems. These problems are not subordinate to the items listed because I can state them so simply and because they are so important in their specificity. Instead, paragraph 3 seems to run counter to these problems: 3.1 Dignity is not about people as subject to AI technology: people are the masters whom AI serves. AI serves to improve human dignity. 3.2 "Freedom" is also what Uber says is the benefit of their platform (never mind the insecurity or unregulated work hours). Freedom is also the freedom of the tech elite to make venture capital decisions, which are also part of the AI ecosystem and should be ethical. 3.4 Platforms can be so monopolistic that fending workers or business off them can be equivalent to denying market access. 3.6 People often find intrinsic value in performing certain functions. They may even be paying a cost if regarded from an economic point of view (poets, writers, artists). These benefits are often not	11. Polder Model. There are benefits and costs beyond money to every job that ever existed that relate to well-being, if there is a more cost effective replacement, then all intrinsic and extrinsic costs and benefits that can be weighed by an A.I. (which may be unfeasible without A.I.) should be taken into account. 12. The context of technological work replacement is systemic, not single apps or startups or individual. Wealth distribution issues, tech elite formation, shortening of the work week should be actively considered and policy should not be blindsided again. Under 2. Technical and Non-Technical Methods to achieve Trustworthy AI - Protocols have to be developed for A.I. communicating with other A.I. in the societal system. - Public administration and industry regulators should seek knowledge partners to experimentally design modules for deep learning neural networks that incorporate ethics derived objective/loss/cost functions. These modules should be paid for by the public and become publicly available, but required by law to be implemented in A.I. applications.	By the EU making available neural network modules that implement ethics guidelines via an open source portal for the EU, organisations can derive implementations for their specific applications (open source versioning) which can be assessed, audited and developed by the public and by auditors. Work psychologists should weigh in.	In scare stories in the media, there is often a reference to the industrialisation of the 18th/19th century, but in the 80's, 90's and 00's and 10's we have seen work change too with paperless offices, outsourcing, online and uberisation respectively. The learning opportunity from these wrt one of the most impactful and all the affiliated and similar consequences of A.I. becoming pervasive is missed by the guidelines: the shortening of the work week. I think that the current guidelines are very much app centric, much more than human centric. Here and there it seems to blue print "cookie consent agreements", missing the wider societal impact. Autonomous driving technology is around the corner. Virtual and augmented reality and ambient computing (the public sphere) are not mentioned. The guidelines seem to play to the defensive, while norms may seem restrictive, the lack of norms made uberisation possible while public administration was blindsided and worried bout whether taxes would be paid. Come to think of it: taxation is missing. The costs of using Google (24USD per query) are relayed. Google famously doesn't pay tax but puts a brain drain on public finance. etc. It's all these societal, systemic, proxy effects that have the force to wash away "cookie consent notices".

				expressed monetarily. Helping people, strength, purpose, social contact, solitude, danger, challenge, power, status, respect, being knowledgable; the instances are myriad and possibly innumerable. AI should strive to implement these considerations into technology. pp. 13 We invite those partaking in the consultation to share their views thereon. Russia has developed undetectable weapons of mass destruction that may strike with no notice and Russia also has a dead hand launch system. Maybe AI should also be used to encertain the continuance of the human race.				
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	
Wim	Aerts				5.Critical concerns raised by AI5.1. Identification without consent. Automatic identification should be restricted to specific private environments where security requires person identification. Informed consent for automatic identification can be obtained for private environments when the person Is granted access to the environment. (E.g. When signing a contract with an employee)In public area’s such informed consent cannot be obtained in a reasonably robust way. Therefor automatic Identification should be restricted to targeted surveillance with a valid legal basis (e.g. Searching terrorists)5.2 Covert AI systemsNo comments, I agree with the text.5.3 Normative &Mass citizen scoring without consent in deviation of fundamental rights.No comments, I agree with the text5.4 Lethal Autonomous Weapon Systems (LAWS)LAWS should be forbidden globally, just like chemical weapons.5.5 potential longer term concernsWhile the development of Artificial Consciousness, artificial moral agents or unsupervised recursively Self-improving Artificial General Intelligence will probably remain unlikely for a long period, the impact of AIsystems that interact directly with the human brain are a concern of the near future.Human behaviour can be influenced by electronic devices that act on the brain. If these systems contain Covert AI systems, they may alter human behaviour in an unwanted way.Remarks on methods to address the requirements for trustworthy AI:Industry is using quality standards and auditing procedures for products en services		

				and production processes that are developed by international bodies like IEEE and ISO.The Commission should take action to encourage these organisations to adhere to the proposed Ethics guidelines for trustworthy AI		
Anonymous	Anonymous	Anonymous		There is this fundamental principle that is shared by many in the realm of ethics : "there is no rights without duties". A world with only rights and no duties would be clearly unbalanced and would probably lead to the disempowerment of the citizen. It has to be noted that in the US the notion of duties is replaced by the notion of responsibilities which come with the citizenship rights. In the Lisbon treaty, the notion of rights and obligations of natural persons is mentioned (art. 7.3) although very incidently and without any details on the obligations. Furthermore, rights shouldn't be for citizen only and obligations for the EU and states only. Therefore, I proposed the following addendum in Part 1 of the chapter 1. Part 1 : It has to be stated that while fundamental human rights are the capstone of EU values, rights granted by EU member-states citizenships come with obligations most of them induced by the rights, principles and values of the EU. In this document, "Rights" should be understand as "Rights and obligations for the citizen, the EU and the states".		One question on Lethal Autonomous Weapons : It is stated in the Lisbon treaty : "that the policy of the Union in accordance with Article 42 shall not prejudice the specific character of the security and defense policy of certain Member States.... ". Therefore, I wonder whether LAWS should be even mentioned in the document as a such policy may be seen as a contradiction with the Lisbon treaty. A jurist insight might be helpful.	
benedikt	herudek	private				the form (large document, convoluted writing) of this request for feedback is not adequate. This text should be split up in smaller, much more readable sections and get supported with video messages. There should be office hours and physical meetings and call ins to discuss. Also, if one would take the effort to go such a large and difficult to read document it would need to be clear, the comments are read, one gets contacted and what the process is to consider comments. This is not a good way to ask for feedback and gives the impression to be rather about making a tick in the box in the participation box rather than being serious.

Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
BERANEGER	Jérôme	ADEL	Rien à redire!	Je suis d'accord avec les 4 principes éthiques mais pas avec celui sur l'explicabilité. Pour moi, le consentement libre et éclairé est associé directement au principe d'Autonomie. Du coup, le principe d'explicabilité serait plus à considérer comme une règle éthique qui découle du principe d'Autonomie, que un principe éthique à part entière !Dès lors, l'encadrement éthique autour du traitement algorithmique peut se découper selon cinq catégories d'éthique interdépendantes les unes avec les autres et qui se complètent : Ethique de la Donnée - Ethique des Systèmes - Ethique de l'Algorithme - Ethique des Pratiques - Ethique des Décisions. Cette évaluation est constituée de 36 critères éthiques répartis selon cinq familles d'éthique qui constituent l'éthique du numérique : Traçabilité - Sécurité - Organisation - Intégrité - Accessibilité - Collecte non sélective - Fiabilité - Protection - Finalité - Biais - Qualité - Explicabilité - Transparence - Autonomisation - Autonomie - Adaptabilité - Cohérence - Automatisation - Confidentialité - Applicabilité - Performance - Culture - Régulation - Déontologie - Trustworthiness - Vie privée - Accountability - Inclusion - Déshumanisation - Autonomie - Libre arbitre - Gestion - Gouvernance - Responsabilité - Annonce - Environnement	Rien à redire!	Rien à redire!	Je viens de rédiger un rapport intitulé : VADE-MECUM SUR LA RESPONSABILITE SOCIETALE DE L'INTELLIGENCE ARTIFICIELLE (IA). VERS UNE IA ETHIQUE ET RESPONSABLE … Préambule 3Chapitre I : L'écosystème numérique de l'IA 81. Cas d'usages de l'IA 102. Environnement digital 13Chapitre II : Questionnement sociétal et moral autour de l'IA 161. Quelle place pour l'homme dans la société numérisée 172. Interrogations d'ordre technologique et sociétal 233. Interrogations d'ordre éthique et moral 30Chapitre III : L'approche éthique relative à l'IA 341. Qu'est-ce-que l'éthique ? 342. Principes éthiques généraux 363. Problématiques et enjeux éthiques spécifiques au digital 424. Critères éthiques et meilleure évaluation des risques des projets digitaux relatifs à l'IA 49Chapitre IV : Le cadre éthique associé e à l'IA 661. Charte éthique autour de l'IA 662. Recommandations relatives à l'IA 733. Régulation associée à l'IA 904. Gouvernance des systèmes algorithmiques et des données numériques 995. Responsabilité algorithmique 111Conclusion 116Je pense que vous devriez rédiger et élaborer une charte éthique sur une IA responsable et humaniste
Pierre	MONGET		To comply with the non-discrimination right of human beings, the AI should be designed and fed with data in accordance with local judicial systems. Indeed, EU countries have different judicial systems with different laws and rules.	Part I.1 How could we insure an AI can understand and respect free will (individuals are free to make their own choices) without applying it to itself ? Part I.4 "Technological transparency implies that AI systems be auditable". It would be interesting to provide guidelines of such IT & procedural audits. Part I.5.1 We need to simplify user consent with a clear list of specific validation (tick boxes) : automatic voice detection, face recognition, ... And explain for each function in which context it will be used (identification + fraud detection, identification + vocal commands ...) Part I.5.1 "Anonymous data". As AI will become more intelligent, it will be more and more difficult then impossible to really anonymise data Part I.5.1 "AI developers and deployers should therefore ensure that humans are made aware of – or able to request and validate the fact that – they interact with an AI identity." The more an AI looks like an human, the lesser the user is prone to interact with. In the future, it must be clear the user is interacting with AI systems to not hinder trustworthy AI acceptance Part I.5.4 "Lethal Autonomous Weapon Systems (LAWS)" Probably the most dangerous AI application,	Part II.1.2 Training AI systems with malicious data sets could lead to AI breaking the principles of beneficence, non-maleficence, autonomy and justice. We need a mean to validate the ethical proof that data fed to self training AI are not malicious Part II.1.4 From autonomy level 4 with LAWS, we must maintain a human driven governance at all time Part II.1.5 Discrimination : the human values used to discriminate demograhics may be biased (based on human nature) thus making it difficult to implement fair values for AI systems Part II.1.9 Safety : it could prove useful to implement a kill switch in AI systems to prevent critical situations Part II.2.2 "non-technical methods / accountability governance" : Could the AI designer & developer team may be named and somehow linked to the AI they have released ? To ensure tracability, accountability and provide a single point of contact to stakeholders	Note : "Use case : Autonomous Driving" If a collision with a group of people is inevitable, what rules are to be implemented in order to preserve the driver or the group's life ? A conflict may rise between the principles of beneficence and non-maleficence	Personal note : I thank the AI HLEG for its initiative of releasing the draft and making it possible for stakeholders to contribute to it. As a EU Citizen, and with engineering experience in different sectors, I am glad to participate in this consultation and I am at your disposal to further discuss on my comments. Best regards, Pierre MONGET

				with big risks of malfunction bypassing human control or selecting wrong targets or becoming out of control. Also, the consequences of LAWS being hacked are catastrophic (change target selection, or bypass clearance to open fire) Part I.5.5 "AI systems that may have a subjective experience of Artificial Moral Agents or of Unsupervised Recursively Self-Improving Artificial General Intelligence (AGI)" There is a risk for unsupervised AI systems to improve temselves and develop "free will" if they are endued with a consciousness (against preservation of human agency) Moreover, there is a need to apply rules to conscious AI (being fair for AI might result in being unfair for human beings)		
Anonymous	Anonymous	Anonymous	I'm not seeing any mention of Institutional Intelligence - the accumulated rules and guidelines that business have for how they conduct business. These represent a significant part of the businesses intellectual property and may drive a significant part of their competitive advantage. A primary goal for any institution in implementing AI is to ensure that the AI is consistent with their existing II. If it isn't then the adoption of the AI solution may significantly change the businesses business model - and thus it's financial viability.I think you need to explain the relationship between II and AI in this section and then explain the relationship between the ethical standards you want to hold the AI implementation to and the ethical standards the existing II is held to. Trying to hold AI to higher ethical standards than the existing II is only going to dissuade organizations from adopting it and adversely impact the profitability of those that do adopt it.Glossary EntryInstitutional Intelligence The accumulated knowledge of an organization, often expressed as a set of rules, guidelines and best practices that serves the purpose of educating new members of the organization and ensuring that their actions within the organization are consistent with the organizations operating principles and regulations. It should be noted that a organizations Institutional Intelligence will be related to the execution of its primary goal – rather than being ethical. Institutional intelligence essentially encodes the organizations way of doing business and its competitive advantage. A key area of AI deployment is going to be to automate (in a piecemeal fashion) various bit of an organizations II. Is it ethical to require an organization to challenge it's ethical outlook as it does so – especially if doing so may make it less competitive on the global stage and so disadvantage its shareholders.It also need to be made clear somewhere in the introduction that this applies to weak AI systems (autonomous data evaluation/processing AI) rather than strong (conscious) AI which would qualify for protection under the human rights act in its own right (and has a whole different set of ethical considerations regrading its development and deployment).Things get implemented because someone think they can make a profit from it – either a direct	Chapter 1Section 2 So companies have an obligation to inform customer when decisions are being made from their data by AI and gives customers the ability to opt out and have the decision made by a human instead? What's the bottom line for AI in this case? When does a linear program simply applying if/then statements cease to be a simple program and become an embedded expert system that is subject to these guidelines?AI, especially neural nets and genetically trained AI isn't particularly good at explaining it's decisions. Often no-one understands how the network of weights and numbers that have been discovered to make the AI perform correctly functions. Thus, unlike human made decisions, it is often impossible to reverse engineer an AI made decision into a series of discrete sub-decisions and considerations that can be translated into a natural language in a meaningful way. AI works much more like a hunch or an instinct than a reasoning/logic engine. Humans can examine the input data and work out why the AI should have made the decision one way or the other, but they can't translate the AIs decision making process into something a normal human would understand. Is there a right to know why a decision was made one way or the other? If there isn't, how do you monitor that the decisions are being made ethically?3.2 Sounds like you are forbidding governments from deploying AI into their anti-terrorist and tax departments. Places that they are already lightly to be deployed. Some mention of common good and society vs the individual might be appropriate.People lie and behave unethically. This also sounds like it would impose significant restrictions on an organizations AIs ability to cross check information that an individual has provided. This in turn will result in the system exhibiting unethical behaviours – rewarding liars, denying services to honest folk and reducing its overall competitive advantage and profitability.3.3 This makes sense when talking about the whole of an AI deployment but not when simply talking about the AI element itself. AIs and AI training is a chaotic process. What you get out will make decisions according to its training, but, no matter how good that training, there is always a chance that a decision will come out that is not entirely compliant with a	AccountabilityThat's nice and nebulous. Describes rectification more than accountability though. The AI itself certainly isn't going to be accountable – it's just a machine. So who will be? Which organization will need to make recompense to rectify the issue? The owner, the operator, the developer, the government agency that licensed/certified it? The text makes no attempt to even raise these issues.Data GovernancePruning data to remove errors is ok. Pruning it to remove decisions made by an organization Instutional Intelligence just because the pruner doesn't think they are fair is much more of a problem as it potentially causes harm to the institution and its shareholders.Note that the order the data is presented to the training engine in and the sequence that the training engine process the fields in each record in can also affect the outcome of the training. If the early data contains significant bias, later training records may not be able to eliminate/compensate for it.Design for AllThis seems much more UI and AI. Perhaps more of a concern should be the identification of those with limited judgemental capacity and the focusing of adds and 'special offers' on them.Non-discriminationHow to protect businesses running ethical European AI against business operating elsewhere in the world on less ethical platforms? How to ensure EU business running ethical AIs are competitive elsewhere in the world? How to detect EU citizens being defrauded/exploited by AI's based outside the EU?Machine learning has also been known to discover false correlations – correlations that are in the data, but which do not reflect the real world. So there is a need to protect against new discriminations – some of which may appear entirely non-sensical to humans (e.g. people who drive a particular type of cars, buy white bread and have neighbours who have one or more dogs are bad loan risks).Even a complete set of data (all available input data) will probably not cover all the possible valid data combinations – 30,000 people might live a in a town, but the possibly number of combinations for even a simple set of 10 fields each with 3 choices is around 3^10 – 30,000,000,000 – possible values. All datasets should therefore be viewed as	AccountabilityIf the AI is purchased/rented from a 3rd party what are their maintenance and support policies? Is there anyone in the organization who actually understands how it works? What can they be held accountable for? What do we end up carrying the can for? How can we tell the difference between bad training and an actual broken AI system.Data GovernanceIs the AI allowed, legally, to do things that we, as human employees, are not? If so, how do we handle maintenance and debugging?PrivacyIf the AI reads it protected data, uses it and then discards it, passing on only the result of its decision, does that have different implications for the GDPR laws? How about if the data is categorized by a subroutine running within it's country of origin?	It seems very idealized in some places. A lot of pipe dream about what could be without to much that seems relevant to the real world.It needs more consideration of complexity and the overlapping chaotic systems that the AI is going to be deployed into.Most development these days is agile - so some thoughts on how to integrate that with the testing. AI systems can also be emergent - 3 or 4 base parts which combine to produce a myriad of complex behaviors that make exhaustive testing prohibitively expensive.You need, perhaps, to more clearly articulate the stakeholders to include customers, users and shareholders (many of which will hold shares indirectly via pension, insurance and banking entities).

financial one or a future socio-economic one. A lot of the ‘common good’ discussion doesn’t seem to fit with this. AI isn’t going to spontaneously appear because someone thinks it’ll be nice or cool or helpful. It’ll be there because businesses think they can make a profit or governments think it will help reduce costs and/or improve social outcomes. For my money this needs to be more apparent in the introduction.

subset of its decision making criteria. It is therefore essential that the AI be backed up with a linear program than evaluates its decisions strictly from the point of view of regulatory compliance before they are actioned. The output from the AI will be a good decision, but it may not be a legal decision – hence the need for post decision review.3.4 Some consideration of how to deal with equality when it comes into conflict with institutional intelligence is required here. If an organization has developed different sets to rules that apply a different sets of, say, socio-economic groups if that allowed to continue within an AI implementation of those rules? Could they get around your guidelines by deploying a different AI for each socio economic group? The effect of forcing all the decisions through the same AI and forbidding it from considering the individuals socio-economic situation would be to significantly reduce the organizations competitive advantage and financial viability – which would significantly impact its shareholders.4. Do no harmSo AI weapon systems are out? Even when the ‘enemy’ have them? Massive disadvantage.To avoid harm, data collected and used for training of AI algorithms must be done in a way that avoids discrimination, manipulation, or negative profiling.Now this is a biggie. Assume that you are working with an organizations existing data to train the AI (which is the way it’s usually done). All the biases and knowledge of their Institution Intelligence will be encoded in that data. So does that mean they can’t use it to train an AI to do business there way? If you’re going to insist that they hide information that could be discriminatory from the AI – ethnicity, age, gender, religion, home address, nationality – then, the resulting AI isn’t going to match the II as it is being forced to ignore many significant factors in its decision making – it will be less profitable, which is a harm to the organization and its shareholders.Autonomy.So businesses and goverments are required to provide/continue programs that provide services using non-AI routes? Simply saying that folks can opt out of AI decision making will often leave then with a choice between that and nothing. Just have a look at the current Australian social welfare program – no choice but to use digital web interfaces and no idea what happens to your data.FairnessBe fair to who? Customers or shareholders? Why should business processes implemented by AI be held to a higher standard than business processes implemented by Humans? AI is just a technology, not a malevolent alien entity. These guides need to make sense if you replace AI with Ouija boards or Tarot decks – it’s simply a technology to automate decision making.ExplicabilityExplicability with some AI systems (especially trained or evolved ones) is extremely difficult. There are known cases of a training processes creating systems that produce the correct results – but we cannot understand how they work. You also get problems with emergent systems where there are hundreds or thousand of inputs to consider and the sequence that they are processed in is significant (I know this from personal experience diagnosing problems on a deployed AI system).While humans may process decisions in a linear branching fashion, trained AI systems process it in a

incomplete.Respect for PrivacyAny thoughts on the AIs duty to society and its duty of care? If it detects illegal activity, is it duty bound to report it? Is it allowed to participate in it? Can it’s records be requested by authorities?Likewise if it detects an possibility that it’s owner/user is likely to commit and act of self harm it is required to ignore their privacy and reach out for help? (Facebook might be an example for this, if it ran mood analysis over each uses posts.)RobustnessIf the AI is operating as a chaotic system, this is basically impossible without a complete dump of the state of the AI at the time the action was taken. Depending on the methodology used to take the dump, even this may not work. A snap shot and playback system can sometimes work – at least up until the point where the defect is identified and fixed. Most embedded AI system won’t have this level of sophistication and recording such data can significantly impact the systems performance (and, in the worst cases, potentially change the results of the AI processing the data).RetrainingAIs get trained with a snapshot of data and then deployed. As time goes by this data will become less and less representative of the social-economic environment that the AI is operating in. It will therefore become necessary to retrain the AI every few years – or every time there is a major shift in the socio-economic environment. An example might be the global financial crash of 2008. An AI trained before it would be making significantly worse decisions after it – potentially to the extent that it would measurably impact the companies profitability.This is also a important step in eliminating long term bias and discrimination as it allows the risk aspect for ethnic and cultural groups to be updated to reflect their actual performance for the last few years as other aspects of their disadvantage are corrected.Fall backFall back needs to be at a business process level rather than an application level. If a problem is found with the AI, the business needs to be able to function without it and to review all of its recent decisions in light of the discovered defect.

				holistic, parallel way that's more akin to a human having a hunch or a gut feeling. Not something that can be easily explained.It's also worth noting that training doesn't, necessarily, get you the best solution. It gets you a solution. Over train the system and it gets worse. You may be able to use genetic principles to refine a trained system, but even then you're only going to get to a local maxima.One point for explicability is the need to record when the system was last trained and the set of training data (which will probably be confidential). Things get even more complex when we get to self learning AI which learns as it goes – it's training data is, essentially, everything it has ever processed. As Microsoft have shown, this can lead to some embarrassing outcomes.5.1The reason we give consent is that we have no real choice – consent or go way. If there was a way to use the service without giving consent, that's what most of us would do.So liquor and tobacco retailers can use AI identification, while a shoe seller cannot?5.2 Auto dialers. In Australia they are starting to use them for telephone scams, using pre-recorded messages and AI to process spoken responses.What strikes me as a significant omission is a simple statement that AI should not be used and developed for illegal activities.		
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Anonymous	Anonymous	Anonymous		I 5.2 Covert AI systems: The relevance of this topic is exemplified by the ever-increasing use of chat bots (in either written or vocal communication) where it is not always obvious for the user that the communication partner is not a human. Including the obligation to identify such non-human communication partners into a regulation might be helpful.	II 1.1 Accountability: This section should also discuss how to handle accountability in the case of really severe wrong decisions, e.g. such that cause the loss of human life (Health, autonomic driving etc.). In general, the accountability should be at least as strong as the accountability of a human for the same activity. II 1.2 Data Governance: The paper states "When data is gathered from human behavior, it may contain misjudgement, errors and mistakes. In large enough data sets these will be diluted since correct actions usually overrun the errors, yet a trace of thereof remains in the data". However, this might be too optimistic as the Microsoft chat bot (Tay) failure has demonstrated that one cannot rely on self-correction due to large enough data. II 2 Non-Technical Methods All AI systems should come with a clear description of their limits, including the areas they are intended for and those, they are not intended for, as well as description of input data that the system cannot properly cope with (e.g. an animal recognition system that has been trained with data on mammals		The paper is a valuable step forward in working out and making explicit ethical aspects of AI.

				might not suit well for identifying insects).			
Uwe	Haass	Roboconsult	Glossary / AI:I suggest to add in the paragraph "As a scientific discipline…":--> knowledge = facts, rules, uncertainties; knowledge from physics, chemistry, etc.; every-day knowledge, domain knowledge, common-sense knowledge etc.;formalised representation and knowledge processing: logic programming, statistics, models,…knowledge acquisition = from observations, from experiments, from books, from drawings,…*Importance* of these comments with respect to Ethics:• Whose knowledge is used/ acquired? Is there an "ownership of knowledge"? I think there are two aspects: something like patents, and something like "protected work" – if someone spends resources to build a knowledge base – this should be protected (similar to data bases) – or should we opt for "Open Science"?--> this is not covered in the paper• Knowledge Acquisition: There are initiatives to derive "biological" knowledge from animals and humans, some are even by means of invasive measurements (electrodes, from brains, nerves, behaviour of humans, and other activities) --> Extreme: Ray Kurzweil: Singularity AI. To what degree should we accept this?--> these types of knowledge acquisition are not covered in the paper.Rationale and Foresight:Heading "Purpose and Target Audience of the Guidelines".I suggest to add that these guidelines are targeted to formalised principles of technology assessment and that the goal is to prepare a structured check-list.The term "technology assessment" is not used.I think the term "formalised" can demonstrate the similarity to the • formalised assessment of drugs, pesticides, food production, etc.• formalised assessment of environmental risks (when building roads, factories, underwater structures, etc.)--> The term "risk analysis" is not mentioned.	1. EU's Rights' Based ApproachI think the paper should go beyond a "human-centric" approach.The paper does not mention (or, not explicitly)• ethics to preserve the environment, the resources, the habitats and eco-systems (and to actively reverse the current situation)• ethics with respect to our children and subsequent generations• ethics with respect to "justice" for a global development of humans (and not only minorities in our EU countries)• ethics with respect to the effects on employment and structure of work• one could also derive ethics to "act now" (rather than do nothing).--> What has that to do with AI?• AI expert systems can amplify the intentions of the owner / builder: to exploit resources or to preserve resources.• AI systems can take away millions of jobs: what are the consequences, how do we keep the society happy and get the "common goods" for everybody?• Energy consumption: example blockchain: By 2020, the Bitcoin network alone could use as much electricity as the entire world does today.I also suggest to include some sort of stewartship which guarantees that none of the AI components comes from suppliers where non-ethical conditions exist.2. From Fundamental rights to Principles and ValuesI suggest that the paper should move from the (currently strong utilitarian) ethics issues into more measurable welfare-economic issues and introduce the Capability Approach.The Capability Approach (CA), developed by Amartya Sen (Nobel Prize) and Martha Nussbaum, is a relatively new paradigm. Different from previous theories of "well-being", the CA allows the analysis and measuring of individual and societal well-being. CA defines a person's well-being in terms of the beings and doings (the "functionings") a person achieves and his/her capability to choose among different combinations of such functionings. The CA gives a framework for the analysis and measurement of: poverty, health care policy, educational justice, participation, evaluation of development projects, assessment of living standards, etc.One of the most popular indices derived from the CA is the Human Development Index HDI (used by the UN). https://en.wikipedia.org/wiki/Human_Development_IndexThe HDI is a summary measure of average achievement in key dimensions of human development. The index is established for all UN member states.--> Why could this be of importance for AI?• CA and HDI are attempts to strip the term "well-being" from general terms which are "relative" to the interpretations and specific political and economic situations and make it comparable – on an international scale.• We may use some of the components and indices for a check list of AI-systems (such as: "is the AI system improving literacy or life expectancy?")• If necessary, we could add indices or components which are	1. Requirements of Trustworthy AII suggest to combine No. 2 (Data Governance) and 7 (Respect for Privacy)--> More importantly, I suggest to add "ownership of knowledge" (see my comments under "Glossary"). More than data! Does the AI system include the knowledge of experienced medical doctors? Does the AI system include the knowledge of "novice" medical doctors? --> 2 issues: one is ownership, the other one is the quality of the knowledge.	Again, I suggest that the procedure (check-list) for assessing an AI system relates to established procedures in technology assessment and takes also into account the criteria which I suggested to add, such as• ethics with respect to preservation of the planet, and our future generations• ethics with respect to knowledge acquisition from humans and animals• ownership of knowledge (= more than data)• quality of AI systems (i.e., quality of knowledge)• fair chances to all humans including in "developing countries"• ethics w.r.t. employment, the type of work and its dependencies, education, choices and capabilities, and cohesion of societies.• develop assessment indices in relation to HDI and procedures as developed in Technology Assessment methods.	Summary:• Add ethics with respect to preservation of the planet, and our future generations (energy consumption, resources,...)• Add ethics with respect to knowledge acquisition from humans and animals • Add fair chances to all humans including in "developing regions and countries" • Add ownership of knowledge (= more than data) • Add quality of AI systems (i.e., quality of knowledge) • Add ethics w.r.t. employment, the type of work and its dependencies, education, choices and capabilities, and cohesion of societies. • Develop assessment indices in relation to HDI (Human Development Index) and procedures as developed in Technology Assessment methods. . Add stewartship to guarantee that none of the suppliers of components of the AI system violates the ethical standards

				relevant to AI systems, such as: “what is the effect of the AI system on the chances to get better jobs?”--> Hence these indices may help to develop and support methods as described in chapter II and III of the paper.			
Asuncion	Lera St.Clair	DNV GL	I like this introduction to the text but miss reference to the importance of existing core values and accountability mechanisms common to companies, from corporate governance to existing assurance systems and risk management practices. Embeddedness of the recommendations presented in this guideline into existing core values, private governance, and accountability mechanisms of companies can be one of the most important criteria for the usability of this guideline by private sector actors. Corporate ethics is one subfield of applied ethics and I think it will be important to consider corporations and private sector actors as also the subject of these ethical guidelines by linking this domain specific ethics code to existing codes. Business transactions of any type are layered and mediated by both private and hybrid accountability mechanisms that self-regulate the behaviour of companies, align business practices with widely held ethical principles (including human rights), and enable societal trust. Assurance providers (accountant firms, auditors, verifiers, certifiers etc..) are guarantors that things do work as they are intended to work. These types of governance systems are commonly referred to as private or hybrid governance and they are widespread across all socio-technical systems. In general, I find the overall treatment of hybrid and private governance unclear and this document could be substantively improved by inviting an expert to mature these aspects to complement the more individual focused part of the guidance.	Chapter one is written from the perspective of the individual, and that is good but partly incomplete. I would like to suggest the consideration of collective responsibility and collective valued, that is the responsibility of groups, or the values (intrinsic and instrumental) that regulate organisations or corporations. Sections 3 and 4 and very well written and important and I am sure the committee has debated extensively its content. I find however, that one key ethical aspect is missing. One of the biggest problems with technologies like AI is that likelihood they will be used to benefit the lives of the haves and leave further behind the have nots or the interests of future generations and the environment. This inequality in benefiting for the potential of AI is partly addressed by the first principle, the principle of “do good”, but the section is too broad, stating only AI could be a tool to bring good to the world, rather than stating how can that be the case. For example stating the need to create incentives to mobilize investments in the public good to prevent the large chunk of AI investments will go towards those who have the ability to pay not those who need it most, or to increase revenues rather than to issues such as addressing the SDGs or climate change. I read that section more as an aspirational point and not a principle that needs to be then executed.	Chapter II does not account for the importance of private governance and the need to leverage existing private governance mechanism to make AI trustworthy. I suggest to revise the list of 10 requirements and add or change those that are related to private governance mechanisms to covey the key role that hybrid and private governance has. For example, the robustness of an AI application can only be determined as such through a process of standardisation and certification or audit by an independent third party. But then these are also key accountability mechanisms, yet not listed in the description of 1 accountability in page 14. We ensure the safety of most systems through standards and verification and assurance. In short, references to hybrid and private governance mechanisms that ensure many of these requirement for most systems is missing. These appear as methods but are presented as means to ensure laymen acceptance. That is a very limited account. Most business to business relations are layered by these technical methods. Also missing is the importance of these assurance mechanisms to enable scalability and foster innovation. It is not immediately clear to me why standardisation is presented as a non-technical method, decoupled from technical methods. In fact there are many different types of standards: design, performance and procedural standards. A very large amount of standards are technical and AI needs all types. In general, I find the overall treatment of hybrid and private governance unclear and this document could be substantively improved by inviting an expert to mature these aspects.	Section III: The confusing treatment of hybrid and private governance mechanisms in the previous sections of the document leads to a poor section on assessing the trustworthiness of AI. Although I agree with the need for circular, continuous assessment process, the section does not distinguish between the very common methods from assurance that can be deployed for the assurance of AI. The assessment section is written with qualitative questions, and these are good for some aspects but insufficient for others and ignore that there may be existing systems such as software standards or methods that could already be put to work. For example, there is a wealth of methods for ensuring the safety of systems. Although the text rightly recognizes that assessing safety is dependent on what particular system is the AI integrated, this is no substitute for a more in-depth analysis of the existing accountability mechanisms and requirements for systems. For example, as software has become increasingly embedded in physical systems, all the work that is done under the label “safety for cyber-physical systems” applies to AI. Also an assessment list seems thin in light of the need for maturing appropriate assurance of AI. A more infomed treatment of hybrid and private governance mechanisms is important not only to improve how to assess the trustworthiness of AI, but to leverage that in depth knowledge and the many available codes for assurance of the reliability, safety and ethical worthiness of many other technologies, as well as to bring this document recommendations closer to existing practices, even if these need to be revised and new assurance methods matured.	I think this is terrific and super important work, but I recommend the team is expanded with some experts on assurance and on hybrid and private governance to integrate these aspects into the document to align its content with existing practices to bring accountability to technology and more generally to private sector activitites.
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential

Dessislava	Fessenko		Comment to the sub-sections "Purpose and Target Audience of the Guidelines" and "Scope of Guidelines": The draft Guidelines appear to promote a mechanism of self-regulation (by the possibility for stakeholders to sing up to the Guidelines). At the same time, the draft appears to leave some latitude of interpretation/application of the requirements promoted by the Guidelines (to that effect the third paragraph of the sub-section "Scope of Guidelines"). Given that the ethical standards and values set out in the draft are by and large common values embedded in the moral and legal systems of all the EU Member States, it appears intuitive and sensible that the requirements promoted in the Guidelines are to the very least minimum standards, and that certain applications of AI may require greater (not lesser/variable degree of) care/higher ethical standards. It my view, it is worth streamlining the language of those two section so that it conveys this messages more clearly and curbs our ambiguities/ possible interpretations as to the type of commitments required by the organizations that would sign up to the Guidelines. This is also important with view to the overall playing field that this Guidelines would set – in European and globally -- so that the Guidelines indeed manage to reinforce a conceptual framework that is progressive and sustainable, rather than susceptible to interpretations/possible exemptions and ultimately able to make itself virtually redundant.	By way of general comments to the entire chapter: - The ethical framework provided in this chapter represents a good starting point for deliberations. However, that framework would likely be more efficient if it conveys a clear message that those ethical standards would equally apply to AI systems and the human beings that design/devise/employ them. The current level of abstraction in this chapter release the humans behind the AI systems from the responsibility of ensuring inception and operation of a trustworthy AI. In my view, it is necessary that the language of the draft Guidelines are amended to clearly set out that the ethical standards promoted equally to the AI systems as ultimate (semi-)autonomous systems, but even more so to the individuals and organisations that set them up, operate them and control them. - In my view, a stronger emphasis should be given to the role and importance of human oversight already in this chapter. Human oversight (ongoing or at least by way of final decision/recast) should be introduced as an ethical standard with respect to the application of AI to situations that may impact physical integrity, health, legal status, access to justices, social mobility, and similar areas of potentially vital importance to a human being. Comment to the sub-section "The Principle of Explicability": The second paragraph of this sub-section proposes a mechanisms whereby individuals and groups may request evidence of the baseline parameters and instructions given as input to the AI decision making. The draft Guidelines in preceding and following sub-sections recognise the importance of correct/representative parameter setting and non-biased (to the extent possible) data. This raises the question, in my view, whether a form of review by groups and organisation, possible with the involvement of independent adjudication bodies, of the parameters set/data used should not be set as a mechanisms for ensuring adequate transparency and accountability of AI systems and their architecture and operations.	Comments to the sub-sections on "Accountability": In order for the Guidelines to set adequate standards for implementation and realization of trustworthy AI, it appears prudent (similarly to the approach taken to data governance) if this subsection is more specific on the types of accountability mechanisms that would be acceptable given the ethical perspective taken in the Guidelines. Please also consider my comment above regarding importance of setting minimum standards and level playing field also with respect to this requirement. Comment to "Governance of AI Autonomy (Human Oversight): Human oversight and possibly ongoing such or at least by way of final decision/recast, should be introduced as a concrete requirement with respect to the application of AI to situations that may impact physical integrity, health, legal status, access to justices, social mobility, and similar areas of potentially vital importance to a human being. Comment to the "Technical and Non-Technical Methods to Achieve Trustworthy AI": The application of the GDPR has proven that compliance by design is not practically possible to be implemented by vast majority of companies in a socially responsible way. Against this background, it appears sensible that more emphasis is put on testing, validating, and IT audits in order to ensure reliable and trustworthy AI.	The Assessment List provided appear generic. Given the complicity of AI systems, should not this check-list be more granular?	
Ansgar	KOENE	University of Nottingham	Page 1 paragraph under the heading "Trustworthy AI": This pragraph reads like AI advertising fluff. It makes grand claimes with little to no hint as to the way in which AI is meant to achieve these things. e.g. how does AI tackle climate change?	Page 12, section 5.3: Citizens need not just the possibility to opt out, but also an ability to challenge/rectify scores they are given. In the case of gig-economy (e.g. Uber drivers) for instance opting out of scoring would not be good for the gig-worker but they do need to be able to challenge wrongfully given negative scores that could affect their future employment. Page 12, section 5.4: Please spell out TEU (presumably EU Treaty?) as this abbreviation was not previously used. Also section 5.4: When considering the application of AI in the theatre of war consideration should also be given to the use of AI in military command and control (e.g. threat analysis) and the need for explain-ability of AI based recommendations that feed into the military decision making process. Page 13 Key Guidance for ensuring ethical purpose, 2nd item: in addition to "employers	Page 16, section 5, end of 2nd paragraph: "unfair competition, such as homogenisation of prices by means of collusion or non-transparent market" this is an issue that needs to be considered, but it seems like a stretch to refer to this as a case of discrimination, unless such unfair competition practices are selectively applied to some people but not others. Page 23, Key Guidance for realising trustworthy AI: (4th item) "deontology charters" - is this a thing? Are there any organisations that have a deontology charter? (additional item) Take into considerations potentials and implications of unintended uses of the AI system and design to minimize potential negative consequences of abusing the system. A lot of the current problems with internet services are related to naive optimism on the part of the designers who had a specific positive use in mind but failed to anticipate the potential for	Page 25, section 3, 4th item: In addition to specifying what definitions of fairness are applicable, "provide the reasoning as to why the chosen definitions of fairness are considered to be appropriate". Page 26, section 8, "accuracy through data usage and control", 3rd item: How is the reliability/correctness of the data guaranteed?	In the Executive summary (page i) the 2nd sentence of the 3rd paragraph: This sentence doesn't work. Such a blanket statement about benefits outweighing risks doesn't make sense, it depends on the area of application. Better to phrase this as "despite potential risks, the benefits of use vs opportunity cost of not using are so great that we can not deny its use but must rather ensure to follow the road that maximises ..." The glossary should include a definition of "technically robust" since this is specified as part of the definition of "Trustworthy AI"

				and employees, or businesses and consumers" add "governments and citizens" to acknowledge information/power asymmetries arising from applications of AI in government services.	abuse of their system.	
Olle	Häggström	Chalmers University of Technology	On p i, l 14-15, I am struck by the asymmetry between the adjectives in "tremendous benefits" and "certain risks" – a contrast that is then made explicit in the very next sentence, saying that "on the whole, AI's benefits outweigh its risks". This is unfounded. I'm not saying the situation is symmetric or that the balance goes the other way. I'm saying we are far from knowing which way the balance goes. There simply isn't any serious study that systematically goes through the various potential benefits and risks, in order to establish that "AI's benefits outweigh its risks". Given the various risks in Chapter I, Section 5 (more on which below), confidently claiming that "AI's benefits outweigh its risks" is preposterous and risks coming across as motivated by ideology rather than by evidence. (In view of this, it may at first sight seem puzzling that everyone advocates continued or increased efforts to develop AI, rather than an across-the-board moratorium on such development. But the reason, of course, is not that we know that "AI's benefits outweigh its risks", but rather that, for a range of societal reasons, a moratorium is utterly unrealistic. It should also be noted that much of the uncertainty regarding the benefits vs risks balance stems from the fact that future AI policies have not yet been written in stone. This actually strengthens the conclusion of the sentence on p i, l 15-16, that "we must ensure to follow the road that maximises the benefits of AI while minimising its risks".)	On p 11, Section 5.2, the second sentence reads "Otherwise, people with the power to control AI are potentially able to manipulate humans on an unprecedented scale". On this, I have two comments. First, the word "people" doesn't ring quite right here, and is better replaced by "organizations" (or the more neutral "agents"). Second, the sentence risks being read as suggesting that as soon as the proposed non-covertness is implemented, there is no such risk of manipulation. That is clearly wrong. People are generally very willing to interact with AI systems in a way that exposes us to manipulation (most of us do that with Google and Facebook every day, and the success in China of Microsoft's Xiaoice chatbot is another example worth studying closely), and the level of manipulation may well be aggravated in the future even in the absence of covert AI systems pretending to be human. Regarding Section 5.3, it should be noted that in China, a large-scale social credit system is well underway. While transparency of any such system is desirable, that in itself does not prevent the system from being used to oppress the population. It should furthermore be noted that formally including an opt-out button in such a system does not guarantee that in practice individuals can opt out without accepting overwhelming costs of various kinds. As to Section 5.4, I wish to emphasize that what is written here – in particular "it can lead to an uncontrollable arms race on a historically unprecedented level" (which, obviously, can in turn increase the risk of World War III) – is on its own a strong indication that the phrase "AI's benefits outweigh its risks" (p i, l 16) that I criticize above is overhasty. Furthermore, I think the claim that "LAWS can reduce collateral damage, e.g. saving selectively children", although factually correct, is nevertheless unfortunate because it risks clouding the fact that a LAWS arms race is overall a bad thing (and you can try and see how the sentence would sound if you replace "children" by the perhaps equally plausible "Christians" or "whites"). In Section 5.5, Footnote 18, it would be worth pointing out explicitly that a development where "self-conscious AI systems would need to be treated as ethical objects" (or should that be "subjects"?) would undermine the human-centered ethical foundations surveyed in Section 3.Staying in Section 5.5, I have two comments on the passage about "Unsupervised Recursively Self-Improving Artificial General Intelligence (AGI) – which today seem[s] to belong in the very distant future". First, concern here should be more generally about superintelligence, for which Unsupervised Recursively Self-Improving AGI is just one of the ways in which it may come about; see Chapter 2 of Nick Bostrom's Superintelligence: Paths, Dangers, Strategies (Oxford University Press, 2014) for a number of others. Second, the statement about belonging "to the very distant future" has very shaky empirical	These timely and well-structured guidelines contain much of value for contributing to putting us on a benign AI trajectory. If the unfounded claim in the Introduction about how "AI's benefits outweigh its risks" is corrected, along with an adjustment for the slight overall tendency towards downplaying risks in Chapter I, Section 5, then my enthusiasm for the document will be wholehearted.

				foundation, and is to some extent contradicted by surveys among AI experts (see, e.g., Dafoe and Russell, 2016, https://www.technologyreview.com/s/602776/yes-we-are-worried-about-the-existential-risk-of-artificial-intelligence/). The conception that nothing very drastic can happen in the near term seems to have arisen to a large extent from individual AI futurologists' conscious or unconscious wish to brand themselves as sane and measured (as opposed to being a mad doomsayer) rather than from solid evidence. Worth reading in this context is Eliezer Yudkowsky's 2017 essay There's No Fire Alarm for Artificial General Intelligence (https://intelligence.org/2017/10/13/fire-alarm/) which lists the three most commonly advocated reasons (A), (B) and (C) for thinking that a superintelligence breakthrough is not near-term, and explains that all three point to circumstances that are likely to still hold shortly before the kind of hard take-off that the author considers plausible. All things considered, we are very uncertain about what is the correct time scale for when (if at all) to expect superintelligence. And here we should not make the tempting mistake of conflating "very uncertain" with "very distant". A final remark regarding Section 5.5 concerns Footnote 21. While it is correct to point out the major difficulties involved in estimating the probability of very rare high-impact events, the statement that for events that have never been observed, "probability of occurrence is not computable using scientific methods" is plain false, and suggests an overly crude and black-and-white view of science. Science is not solely about observing relative frequencies in the past and blindly extrapolating them to the future. For a view that incorporates a much-needed amount of nuance, please see Sections 6.5 (for which my blog post http://haggstrom.blogspot.com/2013/10/no-nonsense-my-reply-to-david-sumpter.html constitutes an early draft) and 8.1 of my book Here Be Dragons: Science, Technology and the Future of Humanity (Oxford University Press, 2016).		
Anonymous	Anonymous	Anonymous	Glossary (and further text): The section on "Bias" reflects a fundamental flaw of the current document and large parts of the public discussion. There is the data-induced bias as disucssed properly here. But, regarding machine learning, "bias" in selection of algorithms and internal is fundamentally important and even a sound theoretical concept, the so called "inductive bias". Inductive bias can not be eliminated and manifests in design decisions on the algorithmic side. In short, if there are no assumptions about the nature of the problem, it is impossible to generalize from a limited amount of data, which however is the goal of machine learning. This fundamentally means that TOGETHER with the data always decisions ("inductive bias") govern the learning process. This should be reflected in the consideration of audibility.	Areas of concern: 5.1-5.4 yes, concerns. Fair treatment in the document 5.5.: no concerns here an general AI - despite the high public attention, there is simply no hint that general AI can be developed. Whatever systems we will have, they will also not occur over night (black swan, singularity) and rather develop incrementally, certainly as long as any compenent of embodiment (e.g. robot, technical system) is involved. Further area of concern: Deception by interested parties, businesses, etc. For the forseable future, AI systems will be used as means in certain business models or for particular goals (e.g. in public administration). Due to lack of expertise, or with a lot expertise, many systems may be used for different types of manipulation, nudging, deception and the shaping of personal or public opinions, which can come as part of (unethical) business models or (unethical) purpusful use e.g. in		Audibility should not target a step-by-step trace-back of internal computations of the algorithms underlying AI. The latter is unrealistic and is neither required for other technical systems, e.g. no engineer can in detail explain how the control loops running in a standard car engine interact and how in detail the outcome occurs. However, audibility could and should be required, apart from the data, also for the algorithmic design decisions ("machine learning bias"), which however requires a higher and more general level of expertise and is mostly ignored in the public discussion and also in this document.	The concept of explicability and "understandability of causality for layman" is not clear to me, given that we here talk about complicated mathematics and highly complicated technological implementations. "Understanding" itself here is more a matter of trust that the experts' implementations follow correct assumptions and make correct use of the underlying technology. It will not be possible (and since long has not been possible any more in other domains of technology) to comprehend and understand in much detail for non-experts.

				manipulation of elections or simply through misunderstanding of the realm and effects of the technology.			
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
			General Comment Considering the short time available, the scope of the document is truly remarkable, appreciating the great complexity of the issues! 1 Regarding the Term "human-centric approach" The draft acknowledges the importance of AI on both an individual and societal level. The term "human-centric approach" misleading suggest a focus on the individual. It should be amended to read "human-societal-centric approach"? Or at least clearly defined to include the societal level. 2 RE Scope: Scope of the Guidelines, and the document in general There are some 'meta' issues that are foundational for any AI ethics guideline and action plan that must be addressed in the text. The primary issue is the need to take a view of what (big) data is, and how it should be classified: as an asset, a resource, a public or common good etc.? Flowing from this, can or should we speak of owning or controlling data; or are concepts such as stewardship appropriate? Data has a particular set of properties that differentiates from physical assets; datasets can be duplicated at near-zero cost, used in multiple ways by different people without diminishing their value, and their value often increases as they are combined with other datasets; it is not a finite resource; the use of a dataset for one purpose does not inhibit the use of the same data for social or non-commercial purposes. Data is the bedrock of AI; the position taken in such discourses has a direct repercussion for AI ethics. 3 Regarding quote: "The Guidelines are addressed to all relevant stakeholders developing, deploying or using AI". It is vital that the essential 'stakeholder' of the public / 'public opinion' be fully integrated throughout the document. The 'big data' of the 'general public' is at the very heart of AI; the opinions of the general publics (advisedly plural? must be integrated. See the EGE calls for a wide-ranging process of public deliberation and lays out a set of fundamental ethical principles to pave the way (EGE document Artificial Intelligence, Robotics and 'Autonomous' Systems 2018).	1. Page 5, 6 etc The draft uses fundamental rights as stepping stone to identify abstract ethical principle and to specify how concrete ethical values can be operationalised in the context of AI. This approach (taking ethics as an academic discipline that is a branch of philosophy) is not coherent with academic thinking. An appropriate approach is to follow that of Nuffield and include respect for human rights as a principle (Nuffield states in their "Ethical principles for data initiatives" http://nuffieldbioethics.org/report/collection-linking-use-data-biomedical-research-health-care/ethical-governance-of-data-initiatives four ethical principles Respect for persons: Respect for human rights: Participation: Accounting for decisions..... 2 RE Chapter I: Key Guidance for Ensuring Ethical Purpose:The reference to common good is much appreciated (page 5) . This line of analysis and action should be expanded. For instance although the concept "third-generation human rights" (referring to group and collective rights) remains unofficial, this concept but surely be integrated in the guidelines in some manner. 3 Pages 5 - 10; listing of principles, rights etc. Although the structure of the draft is well appreciated, issue of participation must take a prominent place , not only appearing in section II on "realising...". It should have a place in Chapter I and be included in principles. A statement Along the lines of "the opinions, expectations and concerns all levels of AI R&D should be determined with the participation of people with morally relevant interests" might be appropriate. The opinions and attitudes of people regarding 'their' data need to be monitored and used as inputs in ethics decisions, e.g. if/when personal information should be used by whom to what end; should 'their' date be sold to private companies /commercialised; or is the opinion widely held that data should be shared in order to serve the public good (with /without consent)? 4. From Fundamental rights to Principles and Values; refs to consent in general in the draft. Ethics considerations of AI must include a sophisticated discourse on informed consent	Chapter II: Key Guidance for Realising Trustworthy AI: The term accountability is indeed very important in AI. The guidelines should also include a focus on responsibility (not only accountability) As the EGE document Artificial Intelligence, Robotics and 'Autonomous' Systems 2018 succinctly states, accountability is ultimately related to human responsibility. The AI/algorithm / big data literature widely contains calls for developers and users of AI to move beyond causal accountability and causal responsibility; there is an expectation that organisations create algorithms / AI that provide the greatest possible benefit to people around the world. Organizations should furthermore not be indifferent to how the models they develop are used, by whom the models are used, and how the benefits of their new analytical services are distributed. An analytical ethics deconstruction of the term 'responsibility' can provide a constructive framework to address such pressures; a detailed treatment of 'responsibility' is available from the author. II. Realising Trustworthy AI; terminology "Governance of AI Autonomy (Human oversight)" Autonomy in the ethically relevant sense of the word can therefore only be attributed to human beingsThe use of the term "AI Autonomy" is confusing and best avoided. EGE document Artificial Intelligence, Robotics and 'Autonomous' Systems 2018. The EGE urge the High-Level Expert Group on AI to take their recommendations into account		
Nicola	Stingelin	University of Basel associate researcher; Member Royal Statistical Society Special Interest Group Ethics					Great work Much still to do... Would have made many more comments time permitting

								that is expanded to consider community assent and community representation. 5 See page 6The statement that the AI HLEG is not the first to use fundamental rights to derive ethical principles and values is a tenuous interpretation of Oviedo. 6 Para 3 Fundamental Rights of Human Beings The confusion between rights and ethics principles detracts from the serious intention of the document. 7 Page 11 section 5.1 Identification without Consent There is connection from an ethics point of view between consent and consideration. It is ill advisable to take a legal contractual approach and apply to consent.		
Kristof	Kloeckner								Overall, the document is a well-balanced discussion of what it takes to achieve trustworthy AI, based on fundamental principles for a democratic society in Europe and leading to concrete guidance for building and deploying AI based applications. It rightly builds on existing frameworks like the Oviedo Convention and GDPR that address aspects of other technologies impacting human lives. A few comments:• I believe one of the main risks of AI is naïve implementation of AI-based applications and inadvertent harm created through overoptimistic use, which can also lead to growing mistrust of AI in general (an exacerbation of the 'trough of disillusionment').o Accountability and transparency are vital to counteract this, as stated in the document.• Users (especially decision makers) need to be educated in the limitations of AI in general and the systems they are using to avoid overly naïve trust in the decisions of AI-based systems. I believe there is a very concrete risk of decision makers abdicating responsibilities to badly understood AI systems. In the public space, this can lead to reduced legitimacy of the public institutions they represent, in particular social services.• To counteract this, significant education and the evolution of standards of practice is necessary. One should also consider paths leading to certification of AI practitioners (like the chartering of engineers in many European countries). • Academia and professional communities share a responsibility for building curricula and disseminating best practices. Only with broad agreement on, and good understanding of, these practices can consistent execution of the guidance in Chapter II be achieved.• Professional societies and communities (like GI, BCS and their subgroups) need to become stakeholders of the process of creating trustworthy AI, their role should be explicitly called out.• While much attention is rightly given to proper selection of data to train learning systems, I believe the quality of domain knowledge bases also needs to be called out. Here also, transparency is a necessary prerequisite, but there needs to be ongoing curation and quality assurance. For this, domain specific stakeholder (groups) are necessary.• There may be a case for open-sourcing 'ground truths' for

						specific domains. With proper governance, this would lead to higher consistency and transparency.• The principle of 'informed consent' is central to how citizens retain agency with respect to AI. However, this requires general education to be able to make the necessary decisions of opting in or out. There is a big risk that people (consumers) will blindly opt in, as we can see in wide-spread attitudes towards social media platforms.• How realistic is the ability to opt out of commercial use of AI for any given individual? How can enterprises be compelled to offer services regardless?• Individuals will need the support of strong advocacy institutions (like consumer protection bureaus) to be able to challenge AI-driven decisions that affect them, like withholding services or charging higher fees. While the paper rightly puts a lot of emphasis on protecting minorities from bias, for any individual there is a significant asymmetry of power when faced with large (sometimes monopolistic) commercial platforms. This could be addressed in the discussion of the 'Insurance Premiums' use case.• The robustness criteria in point 8 of the requirements are vital. In additions, error paths and triggers for human intervention need to be explicitly called out. For instance, users of recommender systems need to understand the rationale for a recommendation and be able to overrule it. It might make sense to write a follow-on document that discusses these criteria and their implementation in an AI lifecycle in more detail, preferably in the context of a major use case. Ultimately, a taxonomy of AI applications and a comprehensive risk model is needed, and could be based on similar work for general applications. • The risks called out on page 18ff are valid, and there is evidence that at least some of these areas are being pursued e.g. by China, so safeguards are needed. However, it should be pointed out that this is not only a risk that comes from state actors, but also commercial entities, especially those pursuing an ad-based business model.
Sam	Smith	medConfidential	The guidelines throughout make many references to the rule of law - for reasons and in ways that we fully support and welcome. However, there is no guidance on what that means in an automated decision making context, beyond the comprehensive Venice Commission checklist for public bodies. Some of it clearly applies, parts of it clearly doesn't, and in areas the nuance of AI/automated decision making mean it should go further. When publishing the Guidelines, the Expert Group should suggest that the Council of Europe workstream on the Rule of Law work with other experts to produce a 'checklist' for rule of law in the context of automated decision making and AI. The Council of Europe is already doing some work on algorithms in the context of the freedom of expression, but that is a different question, and not one that relates directly to the rule of law. The current checklist from the CoE is here: https://www.venice.coe.int/images/SITE%20IMAGES/Publications/Rule_of_Law_Check_L	The checklist mentioned above is useful guidance for technologists who know what the assessment will be.	The checklist mentioned above is most necessary for independent assessment.	

			ist.pdf These guidelines produced by the Expert Group are just that - guidelines by experts. They are not a generally testable or falsifiable checklist, and there is a need for both.				
Anonymous	Anonymous	Anonymous	Glossary: Human-centric AI: if we must ensure human values are the primary concern, we will get nowhere. I can see what is meant, but is it not a well-known problem that human values, the things humans value, can vary quite widely by culture and even within a culture?	3.2: freedom: how far should we carry freedom? There is always the socio-cultural background that will encourage some decisions while discouraging others, should AI be turned toward diminishing such effects as well? Perhaps an apt paper to mention in this context would be Schwartz, B. (2000). Self-determination: The tyranny of freedom. American psychologist, 55(1), 79. 3.3: respect for democracy, justice and rule of law: is it not quite well known by now that companies like Facebook have already interfered with democracy? If the document is not legally binding, is there any other that is that could address that situation? 3.4: equality etc: AI and algorithms in general are eminently suited for applying the same rules to everyone. In this case, the formulation may need to be changed, because it is not so much the rules that are applied that we care about, so much as it is the biased training data that the rules may be based on that lead to undesired outcomes when they are applied blindly. 4 Ethical Principles in the Context of AI and Correlating Values: - Do good: of course very subjective; it is not clear that the material good that is mentioned will suffice, nor that optimizing with only that in mind will give the best overall result. Incidentally, if we really cared about wealth maximization, how come there are many people whose real wages have been virtually stagnant for decades? - Do no harm: especially in the case of psychological harm, this can sometimes take a long time to become apparent. This reduces to a problem that has been plaguing the accelerating pace of technological development for decades: technology develops far quicker than things like psychological and social processes that need to adapt to them. With AI in particular, the development is so rapid that it seems almost hopeless to try to put the brakes on development to see how it affects people before continuing, especially since there are such powerful incentives to keep ploughing on as quickly as possible. - Be fair: striving for equal opportunity in terms of access to education, goods, services and technology: does AI really have a role to play in all of these? If so, perhaps that should be argued more carefully. 5.3: citizen scoring: what happens when the scoring is performed by a private entity, like the German credit rating agency Schufa? This agency apparently disclosed its methods to various authorities and researchers, who were all impressed, but also notes its actual scoring methods have been ruled a business secret worth protecting. This seems to fly in the face of the requirement to be transparent about such an important aspect of life. Schufa apparently has enough clout that if they do not know someone, that person is already considered a significant risk and might only be able to get a loan if	1 Requirements of trustworthy AI (3) “Design for all” sounds like it does aim for a one-size-fits-all solution… Maybe something like “Inclusive design” might serve better? (5) AI can help identify inherent bias: how does it contribute to people’s trust of AI if we use it to identify our own shortcomings? It would seem such identification would require trust as a prerequisite. (6) …overall wellbeing as explicitly defined by the user: do users have time to enter this information? Do they have sufficient knowledge of what leads to their wellbeing? Is it feasible to design systems such that they can be focused on arbitrary wellbeing constraints as specified by users? 2 Methods to achieve trustworthy AI: traceability & auditability: internal and external audits can contribute to acceptance of technology: not a lot of citing on the whole in these guidelines, but here in particular we would like to see some source(s), no? Regulation: may make things more trustworthy on paper, but is worthless without enforcement. Perhaps add some details about that? More discussion will be possible when more detail is provided in the second deliverable...	No more comments.	If inclusiveness/diversity (of backgrounds) is mentioned, it is natural to wonder also what specific kind of diversity is envisioned. There are only so many dimensions of diversity that can be considered, but taking the cross product between them quickly blows up the number of possible backgrounds to consider. It therefore seems sensible to consider which of them is the most relevant for a given scenario. (Perhaps there is research also to help decide this question that investigates to what extent various attributes that contribute to diversity influence relevant aspects of a person’s experience with a system in a given context.) Also, the process may not be about ticking boxes, but if a list of boxes is provided, is that not what usually happens in practice?

they happen to live in a "good" place (see https://www.schufa.de/en/about-us/data-scoring/scoring/scoring-work-schufa/how_does_scoring_work_at_schufa.jsp). That is to say, they might as well be a public service, but they and their methods remain private.
5.5: longer term concerns: this will sound far-fetched because I am inexperienced, but I would like to think there is a grain of truth to be found in it, on careful and/or charitable reading. Very broadly speaking, I am somewhat afraid that AI may be pulling humanity in a direction that we may not want to go. More narrowly, for example, its influence on humanities in the form of the newly arising field of digital humanities may be giving researchers a frame within which certain questions are easier to ask and answer than others, and I don't know that that frame is preferable. Of course this has only ever been the case with new technologies, but it seems one difference at least this time around is just how enthralling AI is proving to be, how far its influence is reaching and how successfully it insinuates itself into so many areas of life. I imagine these very guidelines can be related to a realization that we are indeed headed into a dangerous area; the question is, with this seemingly inevitable progression, will we be able to prevent the various undesired consequences envisioned, or will we be forced to admit they are like an essential part of AI that is almost impossible to weed out once the decision is made to venture into this area? Indeed, can we not see this very process unfold before our eyes? Some parts of the guidelines sound like they should have been written 20 years ago, when something could still be done; especially outside Europe, e.g. in China, things seem to have progressed quite far into a direction these guidelines recommend against (e.g. the thinly veiled critique of China's social credit system). Fatalistically speaking, we have opened Pandora's box, and with so much capitalist incentive, how will we close it again?
Ahem, one other concrete point regarding artificial consciousness would be the following. As far as we are apprehensive about stem cell research and things like the recent claim of a Chinese researcher having successfully modified a baby's DNA, I think we should be apprehensive also about trying to create consciousness artificially. It seems both cases share the problem that we are dealing with things that potentially have their own agency, that can be independent moral objects that we cannot treat the way we treat ordinary test subjects and general objects of experiments. As much as it may be disappointing, as well as difficult to maintain patience, I think patience is imperative; we should first understand what we are dealing with and how we want to treat it before we risk committing grave moral errors.
Perhaps we should be similarly apprehensive also about "upgrading" humans. In general this can get murky very quickly, but in the particular case of wild dreams about upgrading biological minds or "downloading" them into electronic substrates, I think caution is warranted until we have carefully considered the consequences. (We might live forever as an electronic mind, but what if the problem about dying is not so much the fact that your life is finite but that it does not feel

				finished?) One more speculation might be entertained. Again, this is not a new phenomenon, but as technology advances, old skills do get lost and machines take over. The problem is as mentioned above, the pace at which this latest technology seems to advance. Just look at the way people learn programming, the amount of knowledge young professionals have of lower-level languages and how this has changed in just one generation. Perhaps it is good to outsource the writing of boilerplate code to machines, but my problem is that the development is so quick that we barely even get to consider that question; before we can even decide whether we value a certain skill, it has come and gone. Failing all else, one might provide an economic argument why this is an undesirable situation: customers might get stuck with technology that only a handful of people know how to maintain. Perhaps that is an overreaction; perhaps the fragmented world of software libraries is not so heterogeneous yet that one cannot simply learn on the fly what one is dealing with...		
Vivek	Nallur			[Section 4 of Chapter 1 - pages 9 & 10] One of the aspects of "Preserving Human Agency" is the responsibility of AI to reveal all alternative decisions possible. Selective hiding of information/alternative paths may be worse, in some cases, than having no transparency at all since it creates the illusion of having considered all options, when making decisions or recommending action. Another aspect of "Preserving Human Agency" is the sufficient strengthening of the "right to withdrawal". The right to withdrawal is merely notional if by exercising that right, the human loses the ability to achieve a goal that would otherwise be achievable. This alternative pathway to achieving a goal would require stronger guarantees than mere voluntary promise, with no consequences for failure.		
Hiroshi	Nakagawa	RIKEN AIP, Japan		Page 9:AI application system is created by training AI as stated in page 9. In addition to training data, we should care about the input personal data to the resultant AI application system we generate using training data. The input personal data to these AI application systems should be correct, bias-free and up-to-date data.		page 18: 10 Transparency:Is explainability as a form of transparency, as stated at page 18? If an AI explains its internal behavior such as decision making process, it would be very difficult for ordinary people , namely layperson (in this document) to understand. Another way to implement explainability or rather accountability, is to that when a user inputs a question to an AI system asks the explanation about the results, the AI system shows a set of examples of input and results output that are similar to the result in question. These set of examples might help the user understand why he/she gets that results in quite understandable manner, I think, at least for ordinary people.page 19: Architectures for Trustworthy AI:The monitoring process is to be implemented as AI system. In my guess, this monitoring AI is one of the most important application area of AI technologies even if it needs big computer power.page 21: XAIMy question is what stakeholders are supposed to understand AI’s behaviors. It is very hard even for AI experts or AI system developers to understand the behaviors, not to mention for ordinary people who are not necessarily	I do not find an explicit relation between technical and industrial AI development and its regulation. Industries are probably very aware of the effect of this document to them, and the worst case would be their reluctancy to develop AI system or products. Then, it is very important for industry workers to know what kind of risk will occur if they do not comply this guide line as well as what is a desirable AI as stated in this draft. If this kind of suggestion is not deemed as the purpose of this draft, I appreciate it.

						familiar with AI technologies. So, the methods for making understandable explanation is very different stakeholder by stakeholder. Is it necessary to state clearly these methodological differences?page 22: Education and awareness to foster an ethical mind-set:Education is very important but not effective for younger generation, say less than 10 years old or very aged people. For these people who would be user of AI system, AI based assistant systems which help them use AI system properly are really needed instead of education.	
Christine Eve	Gadzikwa	STANDARDS ASSOCIATION OF ZIMBABWE	Thank you for incliding me on the AI HLEG. I have reviewed the draft AI Ethics Guidelines The document is well though out and should generate a lot of interest from stakeholders				Good draft document which captures relevant issues as a whole
Sue	Arundale	FIEC - European Construction Industry Federation				Four particular use cases of Al have been selected based on the input from the 52 Al HLEG experts and the members of the European Al Alliance: (1) Healthcare Diagnose and Treatment, (2) Autonomous Driving/Moving, (3) Insurance Premiums and (4) Profiling and law enforcement. We think that these are all rather narrow and would like to suggest another, related to digital construction (Construction 4.0 or more broadly, "Built Environment 4.0".) This could look at the impact of AI on the broader built environment, which is the place that humans spend most of their time. (NB: FIEC applied to have an expert on the HLG, but our application was turned down. We wonder whether there is an expert on the built environment/construction in the HLG.)	In general, we think that the document is mainly concerned with AI/human relationships, but perhaps the environment should be considered as well.
Cristiano	Fugazza	National Research Council of Italy - Institute for Electromagnetic Sensing of the Environment		The "critical concerns" in section 5 suggest that, especially for assessing compliance with the guidelines by existing AI applications, it is crucial to define which applications shall be considered as AI and which shall not. Just to make a straightforward example, many recommendation systems are based on (narrow) AI systems but it can be very difficult to pinpoint them. Whereas all recommendation systems are expected to comply with GDPR, recognising one of them as AI may pose further requirements.	The diverse requirements that are presented in this chapter make it apparent the need for appropriate metadata (e.g., to "provide in a clear and proactive manner information to stakeholders" as stated on page 23). Is there a specific group/deliverable devoted to this?		
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Urbano	Reviglio	LAST-JD Ph.D.		In few words, I suggest to consider - or at least mention - the principle of serendipity (that is, accidental, unexpected and meaningful encounters of information) in order to avoid potential determinism brought by so-called hyper-personalization of content, particularly in social media. What if we (almost) always receive information we like, we agree, we desire? If algorithms adapt to our hedonistic proclivity, for instance, we may lose our tolerance and decrease our horizons. There is an academic community that is increasingly recognizing that designing for serendipity can help to burst filter bubbles and weaken echo			

				chambers. I argue indeed that we need to seriously consider the necessity to assess and support a degree of pseudo-randomness and personalized serendipity in algorithms in order to maintain human resilience and reinforce human rights on the one hand, and maintain a collective imagination and media pluralism on the other hand. Please, consider the academic paper recently published in the journal Ethics and Information Technology "Serendipity as an emerging design principle of the infosphere: challenges and opportunities": https://link.springer.com/article/10.1007/s10676-018-9496-y			
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Johan	Suykens	KU Leuven, ESAT	- p.i executive summary: "maximises benefits while minimising risks":also mention that it might be needed to incorporate "hard constraints", depending on the particular context. One should also aim at defining the contours/boundaries in which AI can safely develop. - p.iii "assessment list will never be exhaustive": this statement seems to imply that the problem can never be solved. I suggest to rephrase this as"The assessment list can be incrementally/continuously updated"- p.iv glossary: suggestion to correct the following sentence:"Artificial intelligence (AI) refers to systems designed by humans ...":because AI systems can possibly create other AI systems. - Definition of AI (see the other document):the previous definition (of the previous document) was suitable for communication to a broad audience. I would therefore merge the new definition part with the old definition part.I propose to call it then "Extended definition of AI" or Further clarification/complementary explanation", instead of "Updated definition of AI".	- p.6: related to Fig.2: should "good/bad" always be defined in a relative sense with respect to human beings or could it also be defined in an absolute sense? Could one also define "good/bad" at the level of the AI system itself?- p.8: "do good": how to define "good/bad"? (it may e.g. depend on the culture, see e.g. the recent Nature paper The Moral Machine experiment) - p.12 section 5.5: consider mentioning here also:. importance to anticipate (see e.g. talks by Max Tegmark). acceleration effects with the current AI technologies. difficult to predict the future developments. importance to define safe contours for AI development	- p.14: typo in the chapter numbers from here (III -> II) ?- p.18 "reduction of information asymmetry": please clarify- p.19 Fig.3: suggestion to add the following to the figure 3:. define AI system. define goal. define context (currently Fig.3 doesn't add much to Fig.1)- p.21 "Explanation":it would be good to note that requiring explainability may possibly lead to performance loss of the AI system. Depending on the application, besides explainability also other objectives can be important.	- p.26 concerning "bias": one should make a difference here between the AI system and the data. It might be that the AI system in itself is fine, but used in a wrong way, e.g. by presenting a very imbalanced data set to it. The user is rather to blame in this case, not so much the AI system.- p.28 "insurance premiums": in comparison with (1)(2)(4) this is much more specific. Additional explanation is needed here why it is in this list of use cases. I suggest to use a less specific term for (3).- p.28 "never be exhaustive": this statement seems to imply that the problem can never be solved. I suggest to rephrase this as"The assessment list can be incrementally/continuously updated" .	Congratulations already with the first draft guidelines! It is an important first step. Thanks also for offering us the opportunity to give feedback.
Gabriel	Zachmann	University of Bremen			I think, the point on accountability needs more detailed and specific requirements. For instance, I think, AI systems should never take decisions on their own , if human lives are affected in serious ways, and if there is enough time for a human to take the decision (e.g., in jurisdiction, or radiology). Also, it should be clear that it is always one or several humans who are responsible for any decisions taken (e.g., the software developers, or the software company, or the radiologist).		

Alastair	Denniston	University Hospitals Birmingham NHS Foundation Trust, and University of Birmingham	(i) Governance of AI Autonomy (Human oversight): a. We suggest that a standard operating procedure should be defined for cases in which healthcare providers' and the AI's recommendations are not in accordance. These procedures should be transparent and available to both the healthcare provider and patient. The occurrence of discordant decisions should be auditable in order to better understand the limitations of the AI system and potential safety concerns. Moreover, particularly in the case of discordance, it is vital that the decisions of the AI system can be traced and interrogated (i.e. by saliency maps). Therefore, guidance on minimum standards to address the 'black box' issue of AI tools in patient care would be helpful. (ii) Data Governance:a. We suggest that guidance should be offered on the minimal set of information (i.e. regarding patient's demographics, patient flow and the labeling process - in supervised regimen) of datasets used to develop the AI (training, validation and test datasets). We also suggest that this information needs to be accessible to patients, doctors, researchers, policy-makers and regulatory authorities. This will help to assess generalization and potential bias of an AI tool.b. We suggest that guidance should be offered on the minimal requirements of the evidence base an AI tool needs to provide before its implementation in daily medical routine (i.e. regarding the minimal quality of data collection used to develop and AI tool).c. We suggest that guidance should be offered on the minimal level of transparency and accessibility of the datasets used to develop the AI tool.	(iii)Robustness:Reliability and Reproducibility. We suggest addressing the following points:b. Guidance on the appropriate frequency of independent evaluation of an AI system when models are updated with new input data. c. Guidance on testing reproducibility when only the company responsible for producing the AI system has the infrastructure to do so. For example, large companies may produce a system which requires large amounts of computing power which others simple cannot afford. How can we ensure an independent evaluation is carried out on such a system? d. Reproducibility affords transparency/accessibility to the training dataset, access to infrastructure, access to the AI algorithm or the knowledge of its technical specifications.e. The introduction of an AI system to a discipline, including the reporting of its performance level, should be accessible to the target audience. The way in which results are published should respect the context of the discipline. In the healthcare example, diagnostic accuracy metrics (such as sensitivity, specificity, positive/negative predictive value) provide more information than a simple accuracy percentage. Therefore, compliance should be afforded with guidelines established in healthcare research (i.e. on reporting; TRIPOD, or STARD statement). Moreover, this ensures that reviewers can compare clinical utility of AI versus non-AI models. (iv) Accuracy. We suggest addressing the following points: a. We find that currently many AI studies currently do not report a threshold for which the final performance is measured. This is a particularly important metric for diagnostic studies where false-positive and false-negative diagnoses have the potential for patient harm, and therefore the optimum threshold is conventionally set at the point where sensitivity and specificity are the most balanced. Depending on the healthcare context, a higher false-positive/negative rate may be acceptable. If a specific threshold is used when reporting the AI system, this should be explicitly stated.b. In healthcare, false-positive cases identified by the AI may lead to psychological burden for the patient and financial burden for the healthcare system, while false-negative cases may lead to the worsening of prognosis. These "costs" should be assessed, weighted against each other and reported.c. Representativeness: Evidence in the clinical setting and the domain the AI-system will be deployed in should be assessed. (v) Transparency. In our perspective, transparency should address the following points:a. Accessibility to the training datasetb. Access to the AI-building infrastructurec. Access to the AI algorithm or the knowledge of its technical specifications inclusive methods of pre-processing. d. Comprehensive reporting in accordance with published guidelines in healthcare(vi) Traceability.a. Method of testing the algorithmic system. External validation of AI algorithms should be mandatory (to answer the question whether the outcomes are reproducible in settings beyond where the system was developed).We hope that you will consider our comments as a helpful feedback to your draft for ethics guidelines. In case you wish for further information, we would be happy to develop some EU advisory	Following your request for feedback on your draft for ethics guidelines of AI, we would like to submit our comments to the operationalization of the assessment list tailored to the use-case of (1) Healthcare Diagnosis and Treatment. We would like to congratulate the panel on producing this comprehensive draft guideline on responsible AI implementation. We, Prof. Alastair K. Denniston and Dr. Pearse A. Keane, are experts in the fields of AI, digital health and diagnostics in healthcare, and are well-placed to fulfill an advisory role for guidance in this use-case. We are currently in the forefront of the discussion around evaluating AI medical diagnosis and we see you as an important ally in our mission to raise the quality and validity of AI in healthcare. Similar to recent concerns regarding inadequate regulation of medical devices, we find that recent studies validating AI-assisted diagnostic models in medical imaging are not held to equal scrutiny as other clinical predictive models. Whilst many are excited about the potential for AI as a diagnostic tool, we recognize the need to ensure such tools are evidence-based, trustworthy and patient-centric. From our current perspective, key findings are issues of design bias and a lack of consensus on minimum reporting standards which limit interpretation of the evidence around AI utility. Because of poor research methodology and reporting, it is currently difficult for the medical community to evaluate how and to what extent AI systems may add benefit to patient care. We offer the above points specifically relevant for the use-case of 'healthcare diagnosis and treatment'.We hope that you will consider our comments as a helpful feedback to your draft for ethics guidelines. In case you wish for further information, we would be happy to develop some EU advisory role for AI in healthcare.Yours sincerely,Prof Alastair Denniston, PhD MRCP FRCOphthConsultant OphthalmologistUniversity Hospitals Birmingham NHS Foundation Trust& Hon ProfessorUniversity of BirminghamUnited KingdomDr Pearse A. Keane, MD, FRCOphth, MRCSINIHR Clinician Scientist and Honorary Consultant OphthalmologistNIHR BRC at Moorfields Eye Hospital/University College LondonUnited KingdomDr Livia FaesResearch FellowDepartment of Ophthalmology, Cantonal Hospital Lucerne, SwitzerlandMedical Retina Department, Moorfields Eye Hospital NHS Foundation Trust, London United KingdomDr Xiaoxuan LiuDoctoral Research Fellow University Hospitals Birmingham NHS Foundation Trust& University of BirminghamUnited Kingdom

							informed consens: Als Standarteinstellung sollte alles immer auf nein stehen, es darf auch nicht aktiv nachgefragt werden, ob man all dem zustimmt. Dies hat nur zur Folge, dass man einfach einen Hacken setzt und das alles wegklickt. Dies hat nichts mit Informiertheit zu tun. Information und wirkliche Einwilligung kann nur erlangt werden, wenn ohne Zwang die betreffende Person aktiv in seine Profileinstellungen geht, Informationen durchliest und dann von sich aus dem zustimmt. Nochmals, es darf vorher keine aktive Information kommen, die einen dazu zwingt sich damit auseinander zu setzen und vor allem dürfen Feature dann nicht einfach gesperrt werden. Auch ohne die Datensammelwut sind üblicherweise die Features arbeitsfähig oder können so erstellt werden, sie sind dann vielleicht nicht so effektiv aber sie würden funktionieren. Um zu informieren wäre ein legaler weg z.B. eine Pressemitteilungen. Weiterhin sollte kein globaler Konsens möglich sein, um zu verhindern, dass versteckt weitere Angebote eingekleidet werden. Die Praxis zeigt, dass Texte immer länger werden und die Menschen das einfach nicht mehr lesen, eine Art allgemeingültiges Piktogramm für eine gewisse Datennutzung sollte etabliert werden, so dass man mit einem kurzen Blick sieht, was das Unternehmen alles mit den Daten anstellen will. Ähnlich einem cc Modells. Dies verhindert auch heftige Auswüchse, da vorher bestimmte Datenverwendungsarten definiert wurden und es keine bösen Überraschungen gibt. Auch sollte es einen Zwang geben diese Daten und vor Allem alle Auswertungen oder sonstigen Benutzungsmöglichkeiten der Daten in jeglicher Art nur in Europa bleiben darf. Also z.B. auf Servern. Diese Daten und Auswertungen und Erkenntnisse dürfen nicht außerhalb von Europa irgendwo auftauchen. Dies soll verhindern, dass ein Datenmissbrauch und insbesondere auch Firmen Daten nicht an ausländische Geheimdienste weitergeben. Sollte dies bekannt werden, müssten Strafen so empfindlich sein, wie z.B. maximal bis zu 20% des weltweiten Umsatzes, dass Unternehmen kein Interesse haben, diese Daten weiter zu geben, da sie sonst Gefahr laufen in Insolvenz zu gehen. Auch Riesen, wie Google, Apple und Amazon sind heutzutage problemlos ersetzbar durch andere Anbieter, wenn Ihnen ein Marktzugang verwehrt wird (siehe China).
Daniel	Richter	Gigapixel GmbH					
Jan	Broersen	Utrecht University, department of Philosophy		In section 5.5, on page 12, specific feedback is asked concerning potential longer-term concerns. I can imagine that this point raised a lot of discussion, as it is fundamental to the problem of responsible AI. All the other points concern the philosophy of technology in general, and are not so much specific for AI, but the point of AI ever reaching the level of moral agency or consciousness, is. To bring this discussion to a good end, two things are missing, in my opinion: (1) a good definition of what is meant by AI, (2) acknowledgment of the wide range of positions in the philosophy of mind that have a bearing on the outcome of this discussion.			

				For instance, if one believes that AI is necessarily Turing-machine based (thus taking one particular position on what AI is), and one is a physicalist, but not a functionalist, one would see no danger of AI ever acquiring a 'mind' and/or becoming a moral or conscious agent. And this is only 1 of many, many positions one could have. And clearly hundreds of years of discussion in the philosophy of mind did not reach one single conclusion. I would think the long term risks as meant here are unlikely to be real (insofar we are dealing with the kind of AI we have now), but the fact that several philosophers have been able to maintain these views without being conclusively refuted is reason enough to point to the 'possibility' of such risks.		
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Maeva	DERRIEN	Consultant in a private french company				On the Uses case 4 « Profiling » if we talk here about for instance IA in HR department to recruit new employees I think it's important to keep a close look on that topic. For sure it depends of the use case, IA to assess resumes/ CV as to be really equitable but in a way has to follow the same rules as today as not being discriminatory. But when we take the case of postulants that need to answer questions in front of a camera and then the IA will analyze the video to asses the candidates profiles (smile, ton of the voice, etc) it's much more risky. I clearly think the risk is for the company to have same profiles and no diversity in their employees approved by the same IA. Europe needs to have competitive companies with heterogeneous employees. As you will go deeper on this use case for the final version I think it's important to find a way to avoid standardization of profiles chosen by IA and think about diversity to leverage creativity and innovation	First of all, I would like to say thank you for making this kind of consultation and work at an European level. I strongly believe that AI can be a wonderful opportunity if the society embrace the topic and understand the technology. Anyway humankind is and will be behind AI, it's our choice to choose what we want to do about it and Europe has a key role to play. I want to add as a general comment on this document that maybe we miss an important dimension: the environmental impact of IA. In the definition of "Human Centric AI" you stated that "the primary consideration, and forces us to keep in mind that the development and use of AI should not be seen as a means in itself, but with the goal of increasing citizen's well-being." A fundamental and basic way to increase human's well-being for the long term is to protect our ecosystem and stop making our planet poorer. We all know that AI and especially machine learning and deep learning is consuming lots of energies to have strong processing and calculation power. For sure, this consumption will increase in the future as every day we find new uses cases for AI. We need to keep this fast roll out of IA but it can't be regardless of the consequences on the environment. It should be the responsibility of Europe to work to find solutions to have an IA more respectful of the environment. For companies it's not as profitable as finding a new use case so they won't put as much effort on that research. But as soon as solutions will be found the companies should be interested to use it as it will decrease

							their costs. In Europe we have lots of IA talents we need to make them work on that problematic. To do a “Trustworthy AI made in Europe” we need to find a way to make AI as green as possible In my opinion, in that reflection on AI Ethics Guidelines, the environmental protection should be included with guidelines on that topic too.
Anonymous	Anonymous	Anonymous					I support the vision as given by Prof. Erny Gillen in his aricle in the FAZ “Die Ethik-Falle” ofThursday 10.01.2019
Anonymous	Anonymous	Anonymous	You write, that the Guidelines are not an official document of the EC - so what is their intended purpose? You also write "... provide information in a clear manner..." - I'm missing a translation into German	In Chapter I/4 one can read on page 9 ("The Priciple of Beneficence"): AI systems can contribute to wellbeing... You are right: it CAN, but regarding to the development of the "social networks" (ie facebook, WhatsApp etc.) I'm really disenchanted. Too many players will try to get influence to AI systems in a way which will destroy legal and/or democratic systems!		The large number of questions you mention in your Assessment list (pages 24 to 28) shows vividly how delicate and extensive the problems are, when more and more AI systems will be installed.	Do you really think autonomous vehicles can solve the climate problems? What is your proposal to give income to the thousands and thousands of drivers, which will loose their jobs? The hype for AI reminds me to the first years after the discovery of the atomic power: many people thought (and sadly some do until today) that nuclear power plants solves all our energy problems - fading out the riks of accidents and the costs of storage of the atomic waste. The same happens now to AI systems: they will bring "paradise on earth"... Please read "Der Zauberlehrling" (sorcerer's apprentice) one of the most famous lays of J.W. Goethe!
Erny	Gillen		I. Process issues and related risks The credibility and plausibility of the Draft Guidelines strongly depends on their explicability, to use the term introduced by the HLEG-AI as a precondition for trustworthiness (of AI). The Working Document is silent notably about the rationale of the composition of the HLEG-AI, the reasoning behind operated methodological and content choices, the culture and style of internal dialogue processes, the limitations of this consultation, organised in the midst of major holiday breaks (where even the EU-AGM system stands still), … Those pitfalls would — according to the Guidelines — not be accepted if performed by any AI system. But more importantly they do harm to the core intentions of the Guidelines and its accompanying process, as wished in March 2018 by the EGE. I strongly recommend to reconsider the timeline and show true openness for discussing and integrating divergent opinions. Following the EGE the process should foster “a dialogue that focusses on the values around which we want to organise society and the role that technology should play in it”. The final Guidelines should, for instance, clearly demonstrate that they do not primarily serve the interests of those called to be members of the HLEG, especially the many directly involved, and prima facie overrepresented AI companies and their	1) The definition of values and the use of the word “value” throughout the working document lack consistency and clarity! The example given to underpin the ethical purpose circuit is wrong when it comes to the value: the informed consent isn’t a value! The protected value is freedom or self-determination! Footnote 2 refers to values as things which is wrong again. Values are attitudes, inclinations, habits, intentions: they describe concepts, but no things! SPECIFIC AD LAWS Lethal Autonomous Weapon Systems (LAWS) should be banned Under the critical concerns the description omits to refer to many requests by the civil society and researchers to ban lethal autonomous weapon systems. That option should at least be mentioned, if not even promoted by the HLEG-AI! This should, according to my ethical convictions, be the position of the HLEG-AI.	2) Under the critical concerns raised by AI, I suggest to add credit scoring and robotic advise i.e. in the finance industry. 3) The conclusion under ‘Governance of AI’ is inconsistent and dangerous. It is stated that the users preferences and the “overall wellbeing of the user” (which might be contradicting under ethical analyses) should be promoted by systems that are tasked to help users. As this conclusion is about Governance, the HLEG-AI should recall strongly that those preferences should be conditioned by the given laws and rules, standards of arts (in medicine and nursing for instance). I would recommend to delete the last sentence! Respect for Human Autonomy is a key concept throughout the working document. Human faculties can certainly be enhanced, but human’s autonomy should be promoted. Enhancing one’s autonomy from outside contradicts one’s internal autonomy! I recommend to replace “Enhancement” by “Promotion of Human Autonomy” in order to be sound and consistent.	The missing research in Ethics Under the Non-Technical Methods to ensure trustworthy AI, I recommend to prominently add Research in fundamental and applied Ethics in the many fields of AI. Ethics is a philosophical discipline which showed great ability to evolve alongside ever changing environments. There is a great need to identify researchers able to dive deep into the complexity of AI and the complexity of Ethics in order to come up with helpful concepts. There is also a need for the regulators and public administrations to deepen their understanding of modern Ethics as an evolving science to be implicated in the critically needed policy designs for trustworthy AI in Europe. Ethics in AI should not be promoted as an internal and mere technical specialisation, but as a professional, multidisciplinary and philosophical approach in the fields of AI. Ethics in AI, as a term, serves that purpose much better than the wording “AI Ethics”. AI Review Board, ethical reflection and ethics committees In Chapter III, ethical Review Boards are mentioned. In the fields of Ethics in Medicine, IRB’s (Institutional Review Boards) are clearly distinguished from Ethics Committees. Review Boards make sure that the standards of arts are respected and validate certain projects from researchers. Thus they work alongside given rules, whereas (Hospital or National) Ethics	My comments on the Draft Ethics Guidelines, open for consultation, want to contribute to successful and consistent Ethics Guidelines for AI in the sense the EGE asked for in its March, 2018, Statement: The process “should integrate a wide, inclusive and far-reaching societal debate, drawing upon the input of diverse perspectives, where those with different expertise and values can be heard.” The current tone of the Working Document does not properly reflect the potential existential risks of AI as largely perceived by the general public, major scientists and philosophers. Thus, it undermines its intention to promote trustworthiness of AI. The systematics behind the principles is not (yet) consistent and sound and should urgently be addressed before entering into wording and language issues. European Ethical Guidelines for AI should put ethics first and not competitiveness, because Europe has shown and shows that it is able to combine both without giving up the one or the other. Social ethics is more than the sum of individual moral choices; it’s about an ethic of care and solidarity. Consistent and well thought Ethical Guidelines with the ambition to introduce trustworthiness as the North star for AI used in Europe are very much needed and should not be sacrificed under the pressure of lobbyists, short-term political agendas or mere time constraints.

academic consultants, but the European citizens.

I'm painfully aware that it is hard to organising a fair process under time constraints and political deadlines. I, nevertheless, urge the HLEG-AI to reconsider the chosen path, notably also because of the weaknesses in its systematics, as I will demonstrate with the following systemic issues to be addressed first.

II. Eight systemic issues remaining vague, ambiguous or unanswered

1.Who is the moral / legal subject of trustworthy AI?
2.How should AI users be protected?
3.How will a human centric approach (vs. a humane approach) distinguish good and bad intentions, right and wrong actions within a human community composed of people of good will and terrorists?
4. How can be assured that ethics is more than a mere function of and for competitiveness or a risk for AI innovation?
5. How can the EU promote trustworthy AI made in other parts of the one world?
6. Which consistent principles should guide the interaction between human users and AI driven systems?
7. How must the concept of 'informed consent' be designed to serve and to protect users?
8. How could the two guiding lists of principles and requirements be harmonised?

Ad 1: Who is the moral / legal subject of trustworthy AI?

The working document refers to a broad range of subjects while addressing trustworthy AI. Sometimes AI is referred to as the virtual acting subject or the grammatical subject; on other occasions AI seems to be the object of the ethical guidelines. In this case, developers, researchers, producers and even users become the addressees of the guidelines and thus the subject for the trustworthiness of AI.

It would be helpful to clarify this issue right from the beginning.

I would recommend to accept AI partly as the subject of these guidelines from level 3 onwards, following the classification in footnote 24, and to introduce a well thought through concept of shared responsibility for developers, researchers and producers. In this sense AI would be part of a collective subject for which i.e. a special legal body could be created within a new legal framework for autonomous systems (cf. discussions around the Maddy Delvaux proposition in the EU-Parliament).

Ad 2: How should AI users be protected?

The Draft Guidelines should not mix up users (consumers) and producers. This is paramount for the concept of trustworthiness and for the consistency of

Committees deal with the grey zones in individual or policy domains. They provide advise to medical doctors, patients, politicians with good arguments and proposals, but they never decide upon the right or wrong choice. The ultimate choice remains with those responsible to act.

I recommend to use and to adapt the good practices from institutionalised ethical bodies and functions for the fields of Ethics in AI.

the chosen approach, if the HLEG-AI wants to maintain the logic behind the 4 + 1 Principles as inspired by biomedical ethics. Those principles were meant by James Childress and Tom Beauchamp to organise the interaction between the asymmetric competent healthcare professionals on the one hand and the vulnerable patients on the other hand by imposing, according to the tradition of Hippocrates, the burden for the implementation of this specific ethos to the professionals. I'm aware that in some cases the lines between users and producers blur, but that should not happen within the Guidelines.

The EU has a clear role in consumer protection, that should not be given up in the field of AI, especially if the aim is to promote trustworthiness of AI. Mixing up stakeholders is an unacceptable trap, as is the subordination of ethics to competitiveness.

Ad 3: How will a human centric approach (vs. a humane approach) distinguish good and bad intentions, right and wrong actions within a human community composed of people of good will and terrorists?

I completely share the concern that AI should aim at "protecting and benefiting both individuals and the common good". But, the term "human centric approach", as coined in the working document, is strongly misleading. The human community is diverse and many interests are competing with others. But, there are generally accepted red lines about what is bad and wrong. Those boundaries constitute our societies and protect citizens. AI should not serve those members of the human family who, for instance, follow criminal intentions or put at risk citizens or the society as a whole.

In order to semantically avoid the underlying misunderstanding the HLEG-AI could use the concept of an "humane approach" (in the sense of beneficial or good AI) thus, introducing a partly open criterion to discern which humans to serve.

The definition in the glossary partly addresses the expressed concern by saying: "The human-centric approach to AI strives to ensure that human values are always the primary consideration … with the goal of increasing citizen's well-being." If this definition should be maintained I strongly recommend to change "primary" into "main" consideration and to add "in Europe accepted values" before values!

Under the imported Principle of non maleficence the notion of environmental friendliness is introduced out of the blue, thus broadening the scope for responsible AI. The crucial question whether AI should serve Life in general or the common good of the human communities is asked, but remains unanswered. The working document as a whole nevertheless promotes a "human(e) centric approach". I recommend to add environmental friendliness at the beginning of the document as a concern of and for human life, thus including it into an inclusive "humane approach".

Ad 4: How can be assured that ethics is more than a mere function of and for competitiveness or a risk for AI innovation?

In the working document there is a tendency to subordinate ethics to competitiveness. I do agree that ethics can and should foster responsible competitiveness. But, any ethic, worth its name, should not be reduced to simply serve a predefined, but limited ethical purpose, like competition. Ethical reflection can't be domesticated without aborting it, especially within Ethics Guidelines!

I recommend that the normal and healthy tension between competitiveness and ethics should be acknowledged and productively be used for the development of an ever evolving ethical discourse and an evolving discourse about responsible competitiveness. To semantically show this concern it would be worth not to use the term "AI Ethics", but to talk about Ethics in AI, as it is nowadays and frequently done in Medicine, where the standard of art term would be: Ethics in Medicine and no longer medical ethics.

The working document expresses, again and again, scepticism about ethical reflection or ethical interventions. This is absolutely strange for a document which wants to promote ethical guidelines and the document should be cleaned from those jeopardising assertions.

By the same token, some authors of the working document even seem to be convinced that biases are mainly injected into AI and autonomous systems by human designers and testers. They even suggest the primacy of AI (as a subject) to overcome human born biases, as stated i.e. in the Glossary: "AI can help humans to identify their biases, and assist them in making less biases decisions". Trustworthy AI certainly can help to identify biases, but it can also produce biases and overlook others. The ethical discernment should not unilaterally or simplistically be delegated to algorithms, as acknowledged in other parts of the working document.

Ad 5: How can the EU promote trustworthy AI made in other parts of the one world?

Even though I like the "made in Europe" brand and idea, I do not think that the EU can and should limit its ambitions to those AI systems "made at home". The scope of these Guidelines should be AI systems used in Europe, whether made in China or the US. If the EC really wants to promote trustworthy AI, it should envisage to address all systems used on its territories.

As this ambition is clearly mentioned as a goal for the longer run, the HLEG-AI should relinquish the expression "made in Europe".

Ad 6: Which principles should consistently guide the interaction between human users and AI driven systems?

Introducing the four generic principles from the field of ethics in bio-medicine as overarching principles into the fields of AI is

certainly of good pedagogical value and easy to communicate. But, exporting this set of principles necessarily also introduces the invisible line of power balance between AI (as a subject or as part of a collective subject) and the users. The analogy between medicine and patients on the one hand and AI and the users on the other hand does not fully match. More thought and research should be invested in this possible, but limping analogy.

Despite diverse criticisms the four principles have shown that they are able to build a consistent and relatively easy to transmit framework. One of their strengths lies in the presumption that they are comprehensive. The working document, inspired by An Ethical Framework for a Good AI Society, adds a fifth principle which from the perspective of the four Principles, by Childress and Beauchamp, could easily be subsumed under their third Principle of Autonomy. The added Principle of Explicability explicitly refers to the concept of "informed consent" which would be typically a part of the principle of autonomy within the original framework.

In order to be consistent and original (in both senses), I recommend to stay with the four principles and to include the transparency concern (included in the explicability principle) into the third Childress and Beauchamp Principle of autonomy.

The larger problem of explicability in the sense of intelligibility and explainability should be addressed outside of the four comprehensive principles. It best fits as a conditio sine qua non introduction to the set of the four principles, because all four imminently depend on the explicability of AI as an input for ethical consideration, reflection and decision-making. Thus, the fifth Principle should not be part of the closed list of the four principles, but a preliminary principle conditioning the set of the four principles.

Ad 7: How must the concept of 'informed consent' be designed to serve and to protect users?

There are numerous academic and practical discussions around the validity of the concept of 'informed consent' and its meaningful understanding in Medicine and Ethics. Nonetheless it works properly in contexts where embedded into an ethic of care, supporting and promoting the autonomy of the weaker part, while simultaneously excluding dominant or paternalistic behaviour exercised by an asymmetrically more powerful part.

As the Draft Guidelines under scrutiny do not distinguish clearly between the different stakeholders and their (legitime) divergent interests, the introduction of the concept 'informed consent' jeopardises its original intent. It easily becomes the loophole for all kind of strategies of the many stakeholders. The language chosen by the HLEG-AI in the working document goes exactly in the wrong direction: Informed consent shall not be "achieved" but respectfully sought for, if the concept is introduced to protect the user / patient and not, the other way round, the

			producer / medical doctor. Given the obligation of the EU to protect its citizens, this language and possible strategy behind is inadmissible! Users should be protected and not trapped, neither by AI nor by Ethics Guidelines! The HLEG-AI Guidelines must show how they efficiently intend to protect all users and consumers, especially the most vulnerable. Ad 8: How could the two guiding lists of principles and requirements be harmonised? For the reader and user of the Working Document it would be helpful to deal with one integrated systematic approach. Now there are two lists: first the list of the Four Principles from Childress and Beauchamp plus (according to my proposition) the preliminary Principle of explicability in Chapter B I and then "the ten requirements" as "derived from the rights, principles and values of Chapter I" in Chapter II. The Requirements of accountability, robustness and transparency could be subsumed under the preliminary principle of explicability. The Requirements of Governance of AI Autonomy and Data Governance should be listed under the ethical principle of beneficence. Otherwise this guiding principle is completely missing under the requirements! The Requirements of Safety and (the missing) Environmental Friendliness could be subsumed under the Principle of Do not harm. The Requirements of Respect for (& Promotion of) Human Autonomy, the Respect of Privacy and Transparency (in the above mentioned sense) would be well understood under the Principle of Autonomy (from list one). The Requirements of Design for all and Non-Discrimination would be massively enhanced if listed under the Principle of Justice and Fairness, thus avoiding simplistic egalitarian language				
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Anonymous	Anonymous	Anonymous		pp. 10, The Principle of Explicability: "Operate transparently" This is, in my opinion, a VERY important principle. Note that it is also linked to the beginning of pp. 21 (Explanation (XAI research)), as well as to Chapter 3, Assessing Trustworthy AI (pp. 27, Transparency), so I would suggest making brief references from this paragraph to the ones in pp. 21 and 27. pp. 11, 5.2 Covert AI Systems	pp. 21, Explanation (XAI research) See my comment related to Chapter I.	pp. 27, Transparency See my comment related to Chapter I.	

				"A human always has to know if she/he is interacting with a human being or a machine" - Another sentence or two could be helpful here, in terms of explaining the potential collision with the Turing test. I guess I am probably not the only one to notice that. I perfectly do understand the ethical background of this document, but note that another perspective is possible as well.		
Anonymous	Anonymous	Anonymous	there must be always			Labelling must be in place and legally required. Visible or immediately recognisable information must therefore always be provided when AI is involved, so that the consumer recognises this immediately. This should not be a hidden information somewhere at the end of a text, but immediately visible. A missing label should be punished with penalties with a high sums, so that it becomes a legal basis. This will strengthen the rights of the consumer and as a citizen. If it is not a real person who communicates with him (no matter which channels), the real citizen must have the possibility to decide for himself whether he communicates further or whether he rather chooses another form of communication. beside this all contracts based in this non-binding form should be forbidden, this means no AI can make contracts. A requirements should be that there must be a real person who signs the contract with the consumer. Auch einen Zugriff auf Daten von realen Personen sollten über AI Systeme restriktiver behandelt werden oder gar verboten werden. Hier muss der Verbraucher und Bürger regelmässig befragt werden, ob er den Zugriff ermöglicht und falls es AI Systeme sind, so muss er auch wissen, das es algorythmische Systeme sind, die diese Information aus seinen Daten herausfiltern werden.
Anonymous	Anonymous	Anonymous				I consider it essential that the data protection rules in force in Europe are also implemented in the framework of AI and can not be plowed or circumvented by general terms and conditions. We all know from personal experience that nobody is able to read or understand the so-called terms and conditions. Therefore, these are confirmed completely unread. In order to minimize this problem, all companies that want to use personal data should provide the user with a one-page information sheet in understandable form in the local language, in which the key messages are summarized. Furthermore, we absolutely need a clear labeling obligation for AI. If a robot can not be recognized as AI when interacting with a human, it must clearly identify itself as a robot beforehand.

irenee	regnauld	Le Mouton Numérique			Dans la partie Technical and Non-Technical Methods to achieve Trustworthy AI, les méthodes "non techniques" prévoient la consultation des "stakeholders", groupes d'experts ou autre. Quelques questions à ce sujet : • A quel stade de développement du projet les parties-prenantes sont-elles consultées ? Si beaucoup d'argent a déjà été investi, il y a fort à parier que le processus dans son entièreté soit construit pour générer de l'acceptabilité. • Quels projets devront ou non mener de telles consultations (privés ? publics ? Gros ? Petits ? Etc.). • Comment sont « fabriqués » les publics et constitués les panels ? • Qui s'assure de la non partialité du mécanisme ? • Comment s'assure-t-on que les personnes consultées ont le pouvoir de voir leurs recommandations aboutir ? (si on se réfère à d'autres controverses technoscientifiques, le résultat est plutôt décevant - OGM, enfouissement des déchets nucléaires, etc.). Merci		
Abel	Torres	Dataveras	Chapter III: Key Guidance for Assessing Trustworthy AIShould include a Strategy on Trustworthy AI document explaining the overview of policies and guidelines implemented from design till implementation level as well as responsibilities. The Assessment List and other documents should be subjected to the high level view of the Strategy.	5.1 Identification without Consent A too broad formulation of this principle can make many practical applications unfeasible. The core intention is that the potential ID of the system doesn't violate any of the declared principles (a patient can be under automatic surveillance for his own good).5.4 Lethal Autonomous Weapon Systems (LAWS) Europe should lead the effort in regulating LAWS by proposing such document for evaluation at UN.5.5 Potential longer-term concerns In its current formulations creates more confusion than guidance. The focus should be on monitoring future trends and rapid evolution to update the document.One of the main concerns that apparently is not listed is the hacking of autonomous systems (e.g. autonomous cars) which are able to cause direct action and harm over people. Such systems will need additional security regulations.	4. Governance of AI Autonomy (Human oversight)Including the capacity to monitor parameter values or range of values and be notified if a deviation takes place.	- Testing & ValidatingThe performance of AI systems has a high dependency on the training space. The stability outside of the training space should be tested (corner cases) and at run time it should be checked whether input data is within such scope or not.	A good start document but needs more concrete actionable and regulatory approach. Happy to contribute.

João	Martins	Executive summary: - (p4) In public debates on AI I've attended, the focus is put on "trust" usually as a way to dismiss the importance of "explainability"/XAI, which is difficult with methods such as Neural Networks. I'm not sure of the reasons for the choice of this title and focus, considering the PDF document itself is called "Draft AI Ethics Guidelines", for example. -In the sentence "Trustworthy AI will be our north star, since human beings will only be able to confidently and fully reap the benefits of AI if they can trust the technology.", trust is put as a precondition, almost a goal. I suggest this sentence could instead end with, "since human beings will only be able to confidently and fully reap the benefits of AI if it delivers consistent improvements in quality of life or wellbeing". Trust I see as a mean to an end, not as an end in itself. - How are these guidelines compatible with GDPR and the requirements GDPR imposes on data processing? Has it been cross checked? Introduction: Rationale and Foresight of the Guidelines - (p8) The executive summary and document itself put a lot of focus in the "Trust" aspect of AI. A focus on Trust, in the context of the discussions about AI I've been involved with/seen, usually comes as an opposition to a focus on Explainability, and frequently from AI practitioners enamoured with Deep Neural Networks. Here are a couple of examples. o In the exec summary it's said that "Trustworthy AI will be our north star, since human beings will only be able to confidently and fully reap the benefits of AI if they can trust the technology". In my view this has the wrong priorities, it should read "since human beings will only be able to confidently and fully reap the benefits of AI if it delivers consistent improvements in quality of life or well being". Trust is what comes out of delivering that improvement, and not a goal in itself. o In the body of section A it's said "In a context of rapid technological change, we believe it is essential that trust remains the cement of societies, communities, economies and sustainable development. We therefore set Trustworthy AI as our north star". Imagine we were talking about establishing the very first court of law or government in human History. Would you focus on Trust? Or do you focus on accountability, in transparency (the laws, previous rulings), in explaining the rationales? If the court respects those, then yes, it's worthy of Trust. If it doesn't, it's a king imposing divine rulings. o "Trust is a prerequisite for people and societies to develop, deploy and use Artificial Intelligence. Without AI being demonstrably worthy of trust, subversive consequences may ensue and its uptake by citizens and consumers might be hindered, hence undermining the realisation of AI's vast economic and social benefits." - Again the focus on trust as a prerequisite and not as a consequence. If I go to the doctor for the first time, I know s/he's accountable - if s/he delivers wrong diagnostics, I'll be able to complain about him/her, take him/her to	This is a strong chapter, with a positive focus on rights. It does feel as if it wasn't written by the same people who wrote the first pages of the document, however. (p12) "The AI HLEG considers that a rights-based approach to AI ethics brings the additional benefit of limiting regulatory uncertainty" - totally agree with this and suggest Explainability for decisions affecting human lives in critical ways be a Right. Recently there was news about this: https://www.cbsnews.com/news/ai-babysitting-service-predictim-blocked-by-facebook-and-twitter/ Shouldn't someone vetted by this system have the Right to know why it decided something, and demand a correction of the information if wrong? (p12) "Informed consent requires that individuals are given enough information to make an educated decision as to whether or not they will develop, use, or invest in an AI system at experimental or commercial stages (i.e. by ensuring that people are given the opportunity to consent to products or services, they can make choices about their lives and thus their value as humans is protected)" – this focuses on those creating an AI. What about final consumers or people affected by it, shouldn't/couldn't informed consent also apply in those cases? (p14) "3.3 […] AI systems must also embed a commitment to abide by mandatory laws and regulation, and provide for due process by design, meaning a right to a human-centric appeal, review and/or scrutiny of decisions made by AI systems" – this is precisely the point I had been missing from the rest of the document until now. Only think missing is a clear requirement for Explainability. (p15) "These four principles have been updated by that same group to fit the AI context with the inclusion of a fifth principle: the principle of explicability. The AI HLEG believes in the benefits of convergence" – See the inclusion of this as extremely positive and necessary, as it's the first time it's mentioned in the document and – as I said above – trust tends to come as opposite to explainable, in some AI communities. "Given the potential of unknown and unintended consequences of AI, the presence of an internal and external (ethical) expert is advised to accompany the design, development and deployment of AI. Such expert could also raise further awareness of the unique ethical issues that may arise in the coming years." (p15) - this role makes total sense for me. Again I am left with the feeling this part of the document was not written by the same people who wrote the first pages - those mostly focus on technical stakeholders (developers, etc.). I'd also add a reference to this Ethics Expert role, in those initial sections. (p15) "It should also be noted that, in particular situations, tensions may arise between the principles when considered from the point of view of an individual compared with the point of view of society, and vice versa." - It would be useful to have a clarification on what is meant here, and more clear guidance. Saying "bad for a set of individuals but good for society" probably refers to the people who may lose jobs to AI automation, I'm guessing. It's not "particular situations", it may be widespread, as the authors are surely aware. (p15) The Principle of Beneficence: "Do	(p21) 1. Accountability: - The suggestion that "In a case of discrimination, however, an explanation and apology might be at least as important" reads insulting to me. I'd rewrite this section, unless you mean "I'm sorry you're black and the system thought you were a gorilla" (see Google) or "I'm sorry you are a woman and our AI didn't pick you for a job" (see Amazon). (p21) 2. Data Governance: - Many texts make the assumption that if somehow data is cleaned up and has good quality, everything will be all right. But if the data doesn't have biases, it loses the predictive power. E.g., if Women or a minority is more prone to have a certain kind of medical problem, and you remove that information from the dataset, how accurate will your system really get? University of Oxford's "Are all algorithms biased" Futuremakers podcast episode published 21/10/2018 touches on this topic. It's all down to the use made of the tech. If the use is ethical/for good, data has to be representative, and that may mean leaving the biases in. - Data governance is essential (know sources, destinations, access rights), plus Data Lineage when appropriate, to track across sources/systems. GDPR may cover this in part. (p22) 4. Governance of AI Autonomy (Human oversight): - Generally agree with the way this section is written. I'd move the footnote 21 to the main text and include example applications and expected behaviors. - I'd use a word other than "oversight", which the "human looking in" versus "human augmented" view I favor in critical decisions affecting people's lives (medical, education, financial). - "It must be ensured that AI systems continue to behave as intended when feedback signals become sparser" – is this is a complicated way of saying "when faced with rare scenarios for which an AI wasn't trained with enough data"? If so, to the best of my knowledge, this requirement is (next to) impossible to fulfil. This is like asking autonomous cars to behave well when driving underwater or in outer space, or a medical diagnosis system to do accurate diagnoses on people over 110 years old. (p23) 5. Non-Discrimination: - Totally agree with this principle. But solving the bias problem is easier said than done. If this is to be addressed seriously, it could include how that can be done – from getting more data to other methods. The AI-to-detect bias suggestion in the first part of the last sentence sounds too optimistic to me and my knowledge of the industry, I'd either make it more concrete and quote accepted research in a footnote, or remove it. (p23) 6. Respect for (& Enhancement of) Human Autonomy - Why the parenthesis in the title of this section? I'd include the two words. - How do you define "Manipulative Nudging" or "extreme personalization"? (does Amazon do that?) - This section tries to strike a balance between two difficult lines. AI by definition puts people in "non-linear" buckets defined by Neural Networks. That's also where AI use then becomes useful. But any bucket can be seen as a limitation of human	(p31) 1- Accountability - Who is accountable for the sourcing and quality of the data? (p31) 2- Data governance - What mechanisms are in place to track data lineage and transformations? - Is access to data controlled and restricted? - Are there mechanisms in place to guarantee GDPR compliance? (p31) 3- Design for All - How we does the system behave for cases such as people with special needs or in minorities? Are there significant differences in error rates? (p32) 4-. Governing AI autonomy - How fast can human control be exercised? - What mechanisms are in place to appeal about an AI-made decision? (p32) 5- Non-discrimination - Are the significant differences in accuracy when using AI with data of minority citizens, or depending on factors like gender or age? (p32) 6. Respect for Privacy - What mechanisms are in place to guarantee privacy? (p33) 7. Respect for (& Enhancement of) Human Autonomy - If applicable, what Opt-out mechanisms are in place for the affected citizens/users? (p33) 8. Robustness: Resiliency - To what forms of attach is the system NOT vulnerable/is protected against? - How does the system handle un-expected data? - What circuit breakers are there in place? (note: has it happens for algorithmic traders) (p33) 8. Robustness: Reliability and Reproducibility - How are deployed models stored and versioned? (p33) 8. Robustness: Accuracy through data usage and control: - Has the system been tested with data from minorities/gender/age? (p34) 9. Safety - Can the system put people's Safety at risk? If so, what are the possible impacts? (p34) 10. Transparency - (Purpose) Is it clear who or what may be hurt by the product/service? - (traceability/method of building) … Change suggested: Please specify what types of personal data were used AND WHAT THEIR SOURCES ARE. - (traceability/method of building) Is training repeatable? - (traceability/method of building) Is there a way to accommodate possible retraining needs to comply with GDPR requirements?	- As I stated before, while I agree with much of the document and the guidelines, I don't think "Trustworthy AI" is a goal in itself. Just contrast with the just published Smart Dubai's guidelines which explicitly mentions Ethical AI. In some discussions I've attended, the word "Trust" is used as a way to circumvent the fact that sometimes there is no explanation possible by Machine Learning/Data Science systems. Trust is a means to an end (improve people's lives and society with AI), not an end in itself. I doubt the committee will adjust this, but it's something I personally strongly oppose. Note how the document itself is called "Draft AI Ethics guidelines", not "Trustworthy AI" guidelines. - The document includes in a few sessions a set of bullets, principles, rights, etc. All of these make reading the document somewhat of a slow exercise, with frequent need to cross checks and some repetition. Some of the internal inconsistencies are probably caused by this. I suggest annexes/indexes are created with one line summaries of these, for easy reference. A second note is that the contents of the document should be easy to reference, and that's not always the case – I had to resort to subtitles + page numbers, and this will also happen when people are designing systems while trying to follow these guidelines. Not sure if there's a standard EU way of doing this. - One of my key concerns is that, whenever serious decisions are made by AI/automated systems that affect people's lives, it should be clear who is accountable and the decision must be explainable. The document mentions this in a few sections, but in a non-uniform way. There's mention of Explicability, Explainability and XAI, for example, but this is for example omitted from the opening sections. - Another of my key concerns is the Human in the Loop/Human Augmentation scenario, and this is something I feel the document addresses in an incomplete way, focusing more in fully autonomous systems. There is a reason why airplanes, although controlled by an autopilot (arguably an AI system) by more than 95% of the flight time, have redundant humans in the cabin. The same should happen in cases like medical diagnostics, justice, insurance/credit, and others like the ones mentioned in the note mentioned in p35. I would add a clear expectation that for these scenarios, AIs act as a sidekick to a person, and not in the driving seat by themselves. - There is vagueness in quite a few sections of the document. This is easy to understand for a document that tried to be broad, but in some cases I do think more detail could be added.

court, etc. If he consistent delivers quality diagnostics, he's gained my trust henceforth. Getting to a point where trust is gained, there are other obstacles to overcome. But my goal is to be get accurate diagnosis, I don't go to the doctor to gain trust.
• (p9) The document also puts some focus on technological mastery of AI. For example, "(2) it should be technically robust and reliable. Indeed, even with good intentions or purpose, the lack of technological mastery can cause unintentional harm." I assume this means knowing about the Machine Learning algorithms used, what is bias and that techniques exist to address it minimizing loss of predictive power (if at all possible), what is Explainability and what techniques are available to achieve it, etc. But this should in my view have more detail, it should suggest that subject matter experts + someone with ethics know-how is involved in any scenario, and that there are rules to be followed. Even with all the technological mastery in the world, one of these systems can cause harm. Plus, how do you define Mastery? Is Andrew Ng's Machine Learning MOOC enough?
I fully agree with the posture around ethical uses of AI, fundamental rights, diversity, human-centrism. But fundamentally think the first part of the document is itself biased towards the hype of Machine Learning/Data Science, and up until now omits the cross-over with GDPR and what that imposes on Machine Learning systems, not to mention Explainability.

Two more notes on text from the exec summary:
• (p6) "Strive to facilitate the auditability of AI systems, particularly in critical contexts or situations. To the extent possible, design your system to enable tracing individual decisions to your various inputs; data, pre-trained models, etc. Moreover, define explanation methods of the AI system." – these are extremely weak recommendations. "Strive to facilitate", "Define", contrast weakly with the demands made in other bullets, where terms like "Provide...", "Ensure...", "Incorporate…" are used. Also there's no definition of what "critical contexts" are. If it's critical, the recommendation of this document should be that explanation is mandatory. If explanation is not possible, an AI should possibly only be usable in "critical scenarios" (such as credit risk of medical diagnosis) in a case of human augmentation. Also: contrast this with the much stronger way rights are described in points 3.3 or 3.5 further down in the document.
• (p6) "Ensure a specific process for accountability governance" – Again this seems light. If a AI-driven system makes a decision that results in loss of lives or human impairment, who is responsible? The authors who published the algorithm in an article? The implementors of the machine learning library? The data scientists who use the library? The company who provided the data? The regulator who didn't legislate on this? Who goes to jail? I feel the document could delve deeper into the accountability aspect associated with AI, not only when building AI systems, but when issues arise. GDPR doesn't fully cover this. In this part of the document, I'd just include one sentence more giving some more info on what this means.

Good" - Recently I read this article https://sloanreview.mit.edu/article/using-artificial-intelligence-to-promote-diversity/ suggesting that AI is used to overcome biases and make fairer decisions, eventually promoting Diversity and Including. Suggest it as an addition to this paragraph.
(p16) The Principle of Non maleficence: "Do no Harm" - A reference to the later chapter, "5.4 Lethal Autonomous Weapon Systems (LAWS)" would be useful here. A useful reference for me is https://www.reuters.com/article/us-portugal-websummit-un/u-n-s-guterres-urges-ban-on-autonomous-weapons-idUSKCN1NA2HG "It would be "morally repugnant" if the world fails to ban autonomous machines from being able to kill people without human involvement, U.N. Secretary-General Antonio Guterres said on Monday". The final sentence also says: "In either case it is necessary to ensure that the research, development, and use of AI are done with an eye towards environmental awareness". Probably worth an explicit reference to Climate Change in footnote 11, and that renewable energy sources are recommended.
(p16) The Principle of Autonomy: "Preserve Human Agency – the last sentence says: "It is paramount that AI does not undermine the necessity for human responsibility to ensure the protection of fundamental rights." This not being a legal text, I wouldn't expect detail here, but someone will have to address the question: if the use of AI results in an accident where people die, who is responsible? The authors of the technique, the developers, the data scientists, the cloud provider, the decision maker, the testers?
(p17) The Principle of Justice: "Be Fair" - This could have additional clarification. Additionally, I can't fully understand the last sentence, "Humans might benefit from procedures enabling the benchmarking of AI performance with (ethical) expectations." Is this just a suggestion to the industry, a recommendation that said methods are put in place, …?
(p17) The Principle of Explicability: "Operate transparently"
- In my view, highlighting technological and business model transparency and not mentioning data transparency is not enough. I can pick up on a popular AI technique such as XGBoost (which is totally transparent, public, well known), have a business model based on making some kind of decisions with it with differing pricing, and leaving out the data this is pretty generic and useless, not really transparent.
- Again in this section, when it reads "Explicability is a precondition for achieving informed consent from individuals interacting with AI systems and in order to ensure that the principle of explicability and non-maleficence are achieved the requirement of informed consent should be sought.", in my view the word used here should be "must" instead of "should".
(p18) 5.1 Identification without consent
- The position of major cloud players here (including my employer) are possibly relevant here: https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/ and https://9to5google.com/2018/12/13/google-not-selling-face-recognition/

autonomy. The last part of the text tries to set the way to address this, but I'd go further – users should be able to opt-out of data collection/use while still retaining use of the system. Otherwise it's like an app on the phone that won't work if it doesn't have access to the GPS, even if it shouldn't need GPS for anything.
(p24) 8. Robustness
- What is a "secure AI algorithm"? Is logistic regression or a neural network "secure"?
- from a technical perspective, I don't think this is possible/implementable: "Trustworthy AI requires that algorithms are secure, reliable as well as robust enough to deal with errors or inconsistencies during the design, development, execution, deployment and use phase of the AI system, and to adequately cope with erroneous outcomes.". This is like asking a car to protect itself against a human crashing or driving while drunk. Suggest this is rephrased.
- "However, the complexity, non-determinism and opacity of many AI systems, together with sensitivity to training/model building conditions, can make it difficult to reproduce results" – unless there are run-time dependencies of external data, this is actually not true. There are plenty of frameworks that store Machine Learning models as a small file (eg, MLFlow, Azure Machine Learning Services) and can also record inputs and outputs. This is actually simple.
- "Accuracy" and "Resilience to Attack" are missing a "So what?", an action that must be taken as a consequence.
(p25) 9. Safety
- The last sentence, "Moreover, formal mechanisms are needed to measure and guide the adaptability of AI systems." seems out of place in the text of Safety. Suggest clarification/context.
(p25) 10. Transparency
- "Explainability – as a form of transparency" – earlier in the document the "explicability" word is used and detailed. Suggest this is made uniform and use Explainability throughout.
(p25) 2. Technical and Non-Technical Methods to achieve Trustworthy AI
- The list of methods here is good. One concern I've heard in debates on AI is that "AI systems are being developed usually by young males using hammers, but they have never built a house". One thing I think is missing here is the proximity to the issue being addressed, understanding of the Problem Domain. In some cases a purely data-driven approach is enough, but in others (eg, who to run over in case of accident with autonomous driving, what's the impact of a mislabel when looking for signs of cancer in images) people involved in the system should have an awareness of the "real world" issue they are solving and the impacts of the solutions. Otherwise it's just "Blind AI".
(p26) Ethics & Rule of law by design (X-by-design)
- "able to take adversarial data and attacks into account" – suggestion: this can be made as a best effort only. Attacks (such as slightly changing images so that recognition fails or mis-identifies) are emergent and not much as known in terms of techniques to protect. I'm guessing the percentage of systems that incorporate any kind of protection for this is < 1%. It's hard to protect against the unknown – you typically

One more aspect I think is missing as priority in this section of the document is the use of AI in Human Augmentation/Empowerment scenarios. That, for me, is possibly the best way to build trust in AI systems especially in "critical" scenarios. An AI helping a doctor do a diagnosis by processing exam images or the history of analyses, an AI recommending a credit attribution score to a risk auditor by pinpointing key factors, an AI suggesting a teacher how best to work with a trouble student to overcome learning disabilities, or notifying a school/parents when it thinks a children is about to run away from home or drop out from school. The Human + AI scenario is almost absent from the document.

(p10) In B "A Framework for Trustworthy AI" (and the exec summary) it's said that "(II) Realisation of Trustworthy AI. Mere good intentions are not enough. It is important that AI developers, deployers and users also take actions and responsibility to actually implement these principles and values into the technology and its use.". Agree with the first part, but the responsibility doesn't lie alone with these. Business Decision Makers, Law Makers, Subject Matter experts, Ethics experts are critical parts of this. Will Data Scientists or "Kaggle grandmasters" or CEO's read these AI guidelines?

- I think we need to be realistic. If nothing is done, this will be done, even if out of convenience. Our phones and laptops do face recognition already, and I personally am not sure how/where that information is and what will be done with it. From here to some EU states using face recognition on migrant's faces is a small step, or any state doing mass surveillance is a small step.
- Security and crime are areas where this does seem useful, however, but a) what if people are mis-identified and run into problems because of this? No AI system is 100% fool-proof; b) will there be an option of using a service or product at all without consenting to the identification?

(p19) 5.3 Normative and Mass Citizen Scoring:
- Don't understand what is meant with "in /all/ aspects". What is the intended meaning?
- My reading is that this is an indirect references to China's mass "social scoring" mentioned for example here: https://www.sciencealert.com/china-s-dystopian-social-credit-system-science-fiction-black-mirror-mass-surveillance-digital-dictatorship . This kind of reality doesn't make sense to my set of personal values (and I believe, the EU's).
- I believe this section can be made clearer in regards to what is being talked about.
- one last note regarding "especially when used not in accordance with fundamental rights, or when used disproportionately and without a delineated and communicated legitimate purpose" – or when used without a possibility to recourse (remember AI makes mistakes and neural networks can't explain) or human augmentation in the loop.

(p19) 5.4 Lethal Autonomous Weapon Systems (LAWS):
- I found this position realistic: https://www.friendsofeurope.org/publication/banning-autonomous-weapons-europe-must-take-lead . Regrettably, if the US/China/Russia develop said technologies, and the EU doesn't, the EU would be at a disadvantage in case or armed conflict. Not an expert on how to address this, but I have many concerns about this. It's the stuff of video games and sci-fi's "killer robots".
- I support EU Parliament's decision.

(p19) 5.5 Potential longer-term concerns:
- There are plenty of doomsayers and optimists talking about super-intelligence driven by Moore's law. I personally don't see the current generation of Neural Networks/AI techniques as leading to AGI, but it's just an opinion. More relevant are however other concerns:
- Impact of automation on jobs and impact of jobs losses – consulting companies tend of predict more jobs will be created (see Accenture, Mckinsey), I fail to see how that is possible. I personally foresee job losses at scale, hitting the middle class (and not low qualified jobs, which is difference to previous massive societal changes), and as an obvious consequence a resulting increased societal asymmetry. I've been in debates hearing the optimists talk about "people will be free to explore other more creative endeavors", which for me are blindness. AI will automate things, almost sure that is coming, the world will have to find a way to cope. This is a major concern and one I feel this document should address.
- A second point I think the chapter fails to address are the human augmentation will have to protect the data collection process itself.

(p27) Testing & Validating
- I'd also add here the analysis of failure cases. It may be that failures share common characteristics or affect similar populations.
- Ethics experts SME was mentioned prior in the document. Shouldn't it also be mentioned here.

(p28) Explanation (XAI research)
- for me one of the critical points in the document, well addressed. GDPR also touches briefly on this, it would be a good idea to do the reference.

(p28) Regulation
- I struggle to understand the mention of Apology w/out Compensation as a way to redress. This should be an exception and only for cases with low impact on people's lives.

(p28) Standartization
- Not sure I understand what is said here. Is it a suggestion that Standartization should happen? A proposal? As written, it sounds only like a reference to the fact that there are standards and they can be positive.

(p29) Education and awareness to foster an ethical mind-set
- I'd include here also Business/Technical/Law Decision Makers. I don't think responsibility should only fall in developers and designers. Technical people will typically do what they are told (with exceptions) and may not even be aware of far reaching ethical issues. Interestingly, the summary further below also mentions "Managers", omitted here.

scenarios, using AI systems to complement humans. The “human in the loop scenario”. Doctors assisted by AI systems doing medical diagnoses, teachers with AI learning assistants, etc. Doesn’t the group consider this to be relevant?

(p20) KEY GUIDANCE FOR ENSURING ETHICAL PURPOSE:
- Picking up on comments made before: a) regarding asymmetries of power, I’d include governments and citizens; b) although arguably not a “vulnerable group” or a minority, women are frequently discriminated against by AIs, reflecting societal biases. Worth adding it to this core list; c) the alert for “Remain vigilant for areas of critical concern” is in my view vague and not enforceable / not strong enough. What is an area of concern? How big must the negative impact be to be worthy of concern? Is there any duty to report AI errors to a legislator? Who is accountable, who has to remain vigilant? And what does someone do (imagine, a data scientist, or a tester) if he finds a use of the tech s/he considers to be “untrustworthy”?

Sania	Kalitska	Warsaw Technical University	DEFINITION OF AI: MAIN CAPABILITIES AND SCIENTIFIC DISCIPLINES I. The name "artificial intelligence" (AI) is often used in the public dialogue about modern development of technologies, but everyone understands it in a different way. This bothers everyone to develop AI systems. Since the intention of the EC High-Level Expert Group is to create AI deliverables to arrange this problem with their correction according to changes in this area, it seems to me that you need to emphasize two forms of this problem – one for today and another for the future.A. Today, AI systems are the separate blocks: social systems, systems of inference and decision making, which include machines, people and the environment. The principles of combining these blocks have not yet appeared, and the blocks are currently being compiled in an ad hoc manner.Some of the known recent ML successes exist in areas related to human imitative areas of AI, such as computer vision, speech recognition, robotics and games. Then, perhaps, we should just wait for further progress in human imitative areas. Firstly, we are very far from realizing the human aspirations of AI imitative. Secondly, a success in these domains is neither sufficient nor necessary to solve today's important problems.B. I think that the development of human imitative AI will accelerate considerably with the use of "context" concept in a wide range. Context is the inherent human's sense developing from his birth and influencing his behavior, views, etc. Context in AI systems will greatly increase the ability to manage distributed knowledge repositories that are rapidly changing and which are globally incoherent. Such systems have to deal with the best interactions in making quick, distributed decisions. Today, the context is used in a very narrow range. I think that now it is the most complicated problem, but its solution in different cases will really be a leap in human imitation. Therefore I propose to add the blocks:- "Context" by the block "Reasoning decision making" in Figure 1,- "Context analysis" next to "Machine learning", "Robotics", "Reasoning" in Figure 2. II. I suggest a corrected definition of the AI system in section 4.An artificial intelligence (AI) system is a formation that operates in the physical or digital worlds, perceiving their environment, interpreting the collected ordered or unstructured data, making a decision based on the knowledge obtained from this data for the best action to achieve a given goal, including adjusting its behavior to a specific impact on the environment.I would emphasize the difference between three concepts:AI is a field of knowledge;AI application is a program performing a specific task based on AI methods;AI system is a system, containing AI application(s).III. I agree with the definition of a human-centric approach to AI, but not quite. In the whole document this definition emphasizes that the main goal of human-centric AI is "to increase human's well-being". In the light of recent climate changes, caused mainly by the egoistic human's approach to the environment, mismanagement of the usage of its resources, I would like to correct this definition.In my opinion, the aim of AI is to expand human's possibilities to create a sustainable world even sometimes by changing human's habits or even renouncing them in favor of this harmony. All other	5.1. Identification without Consent It seems to me that you need to protect European citizens from identification without consent, carried out by the systems of other countries, e.g., China, by banning their import to Europe.5.5.. Potential longer-term concernsv A) If the human imitation of AI will be developed, such systems may have self-development abilities, which must be under human control, because there may be incentives such as human assessment, competition in which a person can gamble away. Then the next step will be either bringing human under control or even replacing him. ----------Example: Context and AII will try to show in a simple example how I imagine the future human imitation AI based on the context.We will consider a case: a person appeals to you and begs for an urgent loan. You want to help her very much. In this case there are 3 situations:1. You know nothing about this person, you can only have general opinions, based on your own impression;2. Your friend knows this person and have a fairly good opinion about her;3. You know this person from childhood and you have a good opinion about her.In these 3 situations our knowledge and the risk of loss of borrowed money are very different:1. The risk is the most: you can be sure that the borrowed money will be lost;2. The risk decreases in comparison with the situation 1, because you have a friend's opinion (external source of information, including context);3. The least risk - your experience about person in a long-term (a lot of sources, including contexts).Long-term contact with a person allows you to predict how it will behave in certain situations. We have according behavior towards him. In our case we know, that he will give away the borrowed money (risk is very small), therefore we lend this money (we make decision).----------If human imitate AI system is built on the context base than throughout its existence, the system will analyze the context from various sources, changing own behavior accordingly (taking decisions in different situation).When dealing with such an AI, we must keep in mind that if AI can draw out the appropriate contexts from various sources, pre-empathize in a way that allows to make specific decisions, it almost "logically thinks". This is a situation where a human has to make sure that AI does not cross a "red line".v B) In a slightly closer perspective, if the researcher creates a robot without bad intentions, which may adversely affect the adverse situation, it may have catastrophic consequences. It is particularly dangerous to create an AI system that has bad intentions. For example, a group from MIT (Massachusetts Institute of Technology) created an AI "psychopath" by using an algorithm named "Norman". It may just be an algorithm, but if they dumped this thing into one of those awful Boston Dynamics dog bodies, we would only have a matter of minutes before Killbots and Murderoids started trampling our skulls (This comes from Newsweek).I think such AI must be officially prohibited.	8.Robustness; 9. Safety; 10. Transparency It seems to me that attributes robust and reliable are not sufficient characteristics of AI systems (applications?) for technical trust. Each computing system is characterized by a functional specification, i.e., a description of what the system is intended for, and a non-functional specification, i.e., a description of how well the system is supposed to provide its intended service. If the computing system complies with its functional and non-functional specification, then the system provides a proper service; otherwise, the provided service is improper. Computing systems perform their defined tasks in different circumstances. Every computing system has specific weaknesses, and works under the influence of a number of interferences from external surroundings. Therefore we don't have sole non-functional specification (number of attributes) for diverse computing systems to define trust. There are a lot of non-functional specifications attributes affecting trust, for exampleResilience a) ability of the system to provide and maintain an acceptable level of service in the face of various faults and challenges; or b) the persistence of dependability when facing changes.[From Dependability to Resilience. Jean-Claude Laprie. LAAS-CNRS — Université de Toulouse — 7, Avenue Colonel Roche 31077 Toulouse, France. 2008.]Dependability a) the ability to deliver service that can justifiably be trusted, or b) the ability to avoid service failures that are more frequent and more severe than it is acceptable to the user(s). Dependability has own attributes (classic definition) [Saltzer, J. H., D. P. Reed, and D. D. Clark (1981) "End-to-End Arguments in System Design". In: Proceedings of the Second International Conference on Distributed Computing Systems. Paris, France. April 8–10, 1981. IEEE Computer Society, pp. 509-512]: availability - readiness for authorized actions,reliability - continuity of the correct service,safety - absence of catastrophic consequences for the user(s) and the environment,integrity - absence of unauthorized system state alterations,maintainability - ability to undergo modifications and repairs,confidentiality - absence of unauthorized disclosure of information.Some authors added to Dependability classic definition some attributes [EU Project CONNECT (Emergent Connectors for Eternal Software Intensive Networked Systems, 2010 - 2013]:trust - accepted level of codependence between systems,security - ability to protect information and computing systems from unauthorized actions,performance - ability of a system to accomplish its intended services within given non-functional constraints (e.g., time, memory),timeliness - ability of the system to provide a service according to given time requirements,precision - ability to provide the same results under unchanged conditions,accuracy - ability of the system to provide exact results,capacity - ability to hold a certain amount of data. Identified attributes indicate the system's reaction to various external and internal influences.Resiliency, dependability attributes are closely related with fault > errorà> failure model, essential to the	1. Accountability (Page.23) I propose to define this attribute by pointing three main factors, affecting accountability, ipso facto to guarantee a certain level of trust: a). trustee, b) trustworthy leaders, c). human factors.a). Trustee You must have the trustee as an independent body or a person, who has confirmed trust level with his authority. Trustee's attributes are:3rd party -information about the trustee provided by external entities, e.g., recommendations; action -type of operation performed by the trustee;capability -type of access rights granted to the trustee;competence -level of expertise of the trustee;confidence -status of the system when evaluating the trustee;context -complementary to location awareness;history - past behavior of the trustee.b). Trustworthy leaders In a collaborative software development environment often the most important contributor to trust is a leader. Leaders require trust between them and those who they lead to be effective. We learn the attributes that generate trust based on the environments we are exposed to and hone them basing on efforts and importance that we place on these characteristics. Trustworthy leaders attributes (for software development in Agile environments) are:competence - combination of skills and experience each individual brings to an endeavor; truthfulness - knowing and sharing the truth; deception, even when 'harmless' or even beneficial, will reduce the credibility of every statement going forward;act as they think - words, feelings, and beliefs match actions of the leader;integrity - taking responsibility for their actions and work and making sure that the work of other's is attributed correctly; a leader with integrity will link themselves to a set of moral and ethical principles that are known to the team and organization;reliability - doing always what he promises,loyalty - showing loyalty towards organization and others is a prerequisite for receiving trust from others. accountability - recognizing, admitting and accepting responsibility for their own mistakes;just - leader is just to those of their team and to those who are outside of their team; the actions of a just leader are predictable and measured rather than erratic and extreme.c). Human factorsHuman reliability is very important due to the contributions of humans to the resilience of systems and to possible adverse consequences of human errors or oversights, especially when the human is a crucial part of the large socio-technical systems as it is common today. User-centered design and error-tolerant design are just two of many terms used to describe efforts to make technology better suited to operation by humans. Human reliability or human performance or HU can be affected by many factors such as age, state of mind, physical health, attitude, emotions, propensity for certain common mistakes, errors and cognitive biases, etc. As a result there are next attributes of human factors:Limited Working Memory -the mind's short-term memory is the "workbench" for problem solving and decision-making. Limited Attention Resources- the limited ability to concentrate on two or more activities	I have to pay attention to the following statements in two documents, which I would correct in this way:1. In the topic of AI-definition, two aspects of the definition should be taken into account:- for today (extending human capabilities in various areas of life) and - for the far future (people imitate AI based on the context concept).2. In the definition of "human-centric" AI we have to emphasize the need to create a harmonic world even by resignation from human's habits that destroy the environment.3 The technical trust attributes have to be more extensive.

challenges, as well-being, etc. must be solved as far as sustainable world is developed.In my opinion, the aim of AI is to expand human's possibilities to create a sustainable world even sometimes by changing human's habits or even renouncing them in favor of this harmony. All other challenges, as well-being, etc. must be solved as far as sustainable world is developed.IV. It seems to me that attributes robust and reliable are not sufficient characteristics of AI systems (applications?) for technical trust, because they do not reflect the behavior of different systems with specific weaknesses working in specific conditions.

understanding and mastering of the various impairments, which may affect a system. Main definitions in fault > errorà> failure model are:threats - circumstances that have potential to cause the loss or harm,fault - cause of the transition from proper to improper service,error - system (or part of the system) state that generates a failure,failures - the hypothesized cause of the error,accident or mishap - unplanned event or sequence of events which results in human death or injury, damage to property or to the environment,vulnerability - weakness in a computer-based system that may be exploited to cause loss or harm attack - exploitation of the system vulnerability,control - protective measure that reduces the system vulnerability. A fault is active when it produces an error, otherwise it is dormant. An active fault is either an internal fault that has been activated by the computation process or environmental conditions, or an external fault. Error propagation is caused by the computation process: e.g., an error is successively transformed into other errors, by their collaboration of service components or services. A service failure occurs when an error is propagated to the service interface and causes the service delivered by the system to deviate permanently or transiently from the correct service. Because nowadays is a great demand in trust systems, various companies began to use the trust technology. Seagate® company uses DriveTrust™ technology, implementing security on the hard drive itself, to provide a foundation for trusted computing. As another example we have at Intel's Technology for safer computing, Intel® Trusted Execution Technology (Intel® TXT), defines platform-level enhancements, that provide the buildings block for creating trusted platforms. I presented this material in such a wide range to show the relation between attributes. If it will be used, it will be nice to me. I can show graphically these relations and their dependencies on errors, if necessary. There are also attributes that characterize human influence on the technical attributes of systems. It is a large area, which is why I was not concentrating on it.

challenges the ability to process information needed to solve problems.Mind-Set - people tend to focus more on what they want to accomplish (a goal) and less on what needs to be avoided because human beings are primarily goal-oriented by nature. As such, people tend to “see” only what the mind expects, or wants, to see. Difficulty Seeing One's Own Error - individuals, especially when working alone, are particularly susceptible to miss errors.Limited Perspective - humans cannot see all that there is to see. The inability of the human mind to perceive all facts pertinent to a decision challenges problem-solving.Susceptibility To Emotional/Social Factors - anger and embarrassment adversely influence team and individual performance.Fatigue - physical, emotional, and mental fatigue can lead to error and poor judgment.Presenteeism -some employees will be in the need to belong to the workplace despite a diminished capacity to perform their jobs due to illness or injury.Stress stress can accumulate and overpower a person, thus becoming detrimental to performance. Avoidance of Mental Strain - humans are reluctant to engage in lengthy concentrated thinking, as it requires high levels of attention for extended periods. People tend to overestimate their ability to maintain control by working. The common characteristics of human nature addressed below are especially accentuated when task is performed in a complex work environment. The mental biases, or shortcuts, often used to reduce mental effort and expedite decision-making include:Assumptions - a condition taken for granted or accepted as true without verification of the facts;Habit - an unconscious pattern of behavior acquired through frequent repetition;Confirmation bias - the reluctance to abandon a current solution;Similarity bias -the tendency to recall solutions from situations that appear similar;Frequency bias -a gamble that a frequently used solution will work;Availability bias -the tendency to settle on solutions or courses of action that readily come to mind. I want to point out that comparing the attributes of Human factor with now received AI parameters shows that in most cases, AI exceeds human intelligence.8.Robustness; 9. Safety; 10. Transparency The indicated attributes are included to the attributes of a higher level affecting the trust. Attributes hierarchy and associations are briefly described before (Chapter II). The practical evaluation of these parameters starts with the analysis of the environment that has a negative impact on the system and measures to counteract it in the system. Under these conditions is performed attributes assessment.

Anonymous	Anonymous	Anonymous	"A mechanism will be put in place that enables all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis." Well how binding will this endorsement be in reality? As you state yourself: "The Guidelines are not an official document from the European Commission and are not legally binding." What I saw and still see is that especially companies agree to certain guidelines on the outside, hang up fancy posters promoting these guidelines, but don't really act according to these guidelines in daily company life and don't even try to, at least partly (gender inequality is still a big topic in some companies nowadays despite what is stated on the outside...). I personally want definite laws that ensure the well-being of citizens and effectively prevent harm from AI related risks. As for: "...AI does not cause unintentional harm..." I'm also interested to know in which way a programmer can be held responsible for his code. I actually expect that an AI programmer should have such an expertise that he can embed exact boundaries in his code that define what the program may do or not do in order not to "get out of hand". One more point that is not mentioned specifically is the accessibility of AI for disabled people and providing an understanding of it in an easy form, accompanying documents/instructions in "Simple English", braille, audio guides for the blind, subtitled videos etc. Furthermore, AI in public life (health system etc.) should not come as an exclusive service that only the wealthier citizens can access, everybody should be able to profit from it if desired and tailored medicine should not only be available for people who can afford a private insurance or similar.	Some of my points mentioned in the previous field were mentioned in this chapter so that is already positive. "systematically be offered to express opt out" I would prefer systematic offers to express opt IN. "an internal and external (ethical) expert is advised" What qualifications should such an expert have? In my opinion this should be a person that is not influenced by any religious beliefs but is philosophically "neutral", and in absolute accordance with the human rights. In general, any purely religious/irrational beliefs that are not in accordance with the human rights and/or science/scientific methods/established facts should never find their way into AI code - in my personal opinion this has a dangerous potential (and fits your paragraph "do no harm" quite well I think). "Transparency is key to building and maintaining citizen's trust in the developers of AI systems and AI systems themselves." How can a balance between transparency and vulnerability (prevention from hacking...) be found? "lie detection, personality assessment through micro expressions, automatic voice detection" How do we ensure that the AI really does a good job in these fields? What if an AI, because it is not as far advanced as it may seem, falsely identifies someone by his/her voice or detects a lie where there is none but, let's say, mere excitement? This is an area where I would be very very careful in trusting an AI and would be afraid of false accusations. As for AGIs I wonder - without being an expert - if the same rules that should apply for any AI code can prevent harm in AGIs: define boundaries in the code right from the start, define the harms that may not be done to whom, define the resources that may and may not be used etc. and that the AGI may not change these boundaries under any circumstances. I wonder if setting such boundaries right from the start could effectively prevent harm or if there still could be the danger that the AGI simply "decides" to disregard all of that at one point becaue of a purpose the AGI regards as more worthy, according to its own standards. I wonder if anyone today, expert or not, could actually realistically say if that could be the case or not.	"Discrimination in an AI context can occur unintentionally due to, for example, problems with data such as bias, incompleteness..." Won't this be always the case in the "starting phase" of an AI, especially regarding the point "incompleteness" - when it is still in the "collecting phase" and does not have enough data sets to be able to calculate reliable results (due to not enough representative data). Will there be a "collecting period" in which the AI will merely collect data and its results will not be regarded as valid or will the results be already used but with especially careful monitoring by humans? Furthermore, what I wonder about is how an AI does not fall in the trap of finding false correlations as it sometimes happens in scientific studies? It's quite hard to imagine that this wouldn't happen to an AI or even more so since at this stage you don't have a conscious AI yet that questions its found correlations. Or can something of that sort be built into the algorithm? "When correlation X = more A equals more B has been found check alternative factors YZ... and evaluate if this could be the real influence on result A and check if B is merely a side effect of the factors YZ..." "Systems that are tasked to help the user, must provide explicit support to the user to promote her/his own preferences" Yes, and as preferences clearly change over time, with aging etc. the preferences should be regularly checked by the AI or the human should of course constantly be able to express that former preferences have changed and wishes to get different recommendations etc. "In some cases this can mean that the AI system switches from statistical to rule-based procedure" Shouldn't certain rules (and boundaries) always be working during the procedure to prevent harm or which rules do you refer to, that is not quite clear to me in this sentence. "include the appointment of a person in charge of ethics issues as they relate to AI" - "This can be in addition to, but cannot replace, legal oversight" This is a very important aspect. The appointed person should maybe be controlled by the person that is taking care of legal oversight. It should also be made sure that in any case of ethic advice, fundamental rights must always be kept in mind and the person giving advice should not be influenced (as mentioned before) by religious, irrational, unscientific beliefs which could lead to favoring of certain groups OR neglect of minorities.	Potential grey areas/difficult areas could be weighing freedom of choice against well-being. If one goal is the well-being of citizens how far should AI systems (especially in the health sector) go in allowing or not allowing self-harming behaviour or potential self-harming behaviour? Or a bit more mildly: how far should it go in manipulating humans towards a life of well-being or not, in CONTRAST to what the human wishes? To take a harmless example, let's imagine a therapeutic app that recommends music. The patient is depressed and wishes to chose melancholic music that puts him/her in an even worse mood because his/her disease is influencing him/her. Would it be acceptable for the app to refuse the desired music in order to protect the human's well being or would it be acceptable that the app strongly recommends other music instead or manipulates the patient into chosing the music that will improve his/her mood more than the one he/she would have chosen? And should it be okay that an AI recognizes people with special needs / mental difficulties and sort of pays special attention to their choices and guides them a bit more towards "reasonable" choices, that are good for their well-being? On one hand you could say such an AI patronizes people with mental disabilities, one the other hand you could say the AI helps such people to unintentionally harm themselves. What would be a good balance between the two? You listed four particular use cases of AI. What about the use for military purposes? Could this be included and discussed unter point (4) or should it be kept out of public discussion?	When commenting the above fields I often had my questions or concerns addressed in the next section/chapter but I left the comments as they are to show which concerns spontaneously came to my mind first. I only learned about the possibility to comment yesterday through a facebook post - it is a pity that it seems not many people heard of it before and that the deadline to comment is already the 18th of January. I hope with later drafts - if there will be one before the final version in March - there will be the possiblity to comment once again and maybe the existence of this draft and the possibility to comment on it could be known better / be made more public within the European media. I know my comments are not very well structured but I hope here and there you will maybe find a new thought/concern that has not been directly addressed so far - although, as mentioned, I often had my questions answered in your draft later, in the following section. My main concern/wish is that AI content programming will stay free from unscientific/religious/irrational influence and that a balance will be found between the "right to self-harm"/freedom of choice and well-being for the human person as a goal. Still not sure myself how patronizing a AI should be allowed to be in order to prevent harm, intentional or not intentional. Curious about your next draft and thank you for making me think through this interesting topic. Best regards, Susanne Desic, Germany
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential

Piero	Poccianti	Associazione Italiana per l'Intelligenza Artificiale	We would like to introduce some general issues that might be relevant to understand the impact of Artificial Intelligence and to direct it for humanity’s well-being.We would like to underscore one point: it is hard to speak about ethics of Artificial Intelligence without considering and analyzing the context in which humankind is living today.Let us observe some of the current difficulties in many aspects of the human living system: democracy crisisinequality growth ecological crisiseconomic crisisWhy?The Economist Intelligent Unit claims people disillusion about formal politics all over the world (http://www.eiu.com/topic/democracy-index). Something has to change and progress in AI can be a very important vehicle in this evolution. Clearly AI might be also a dangerous vehicle for people control, influence and repression, from here some of the many worries in the current debate.Furthermore, inequality is growing. In particolar inequality inside countries. Credit Suisse 2017 reported that “The globe’s richest 1% own half the world’s wealth, according to a new report highlighting the growing gap between the super-rich and everyone else”.It is inevitable to generate some key questions: What do we aim at by using AI? What are our goals? It seems quite difficult to speak about Ethics of AI without defining the goals we have.Many researchers from different disciplines claim that we have to change our economic model and our relationship with the Environment. For example, the growth of GDP can no longer be the only index of success for a country.Today, most countries in the world trust in market economy. In this kind of economy, profit and GDP are the main indicators for well-being, while production costs are the only measure of cost. Are we going to measure the good and bad of AI using the same reference system?Artificial Intelligence is a very powerful tool (or, to be more precise, a set of very powerful tools) and many of us believe it will transform our society in a deep way. It is very likely that, analogously to other powerful tools, AI will contribute to achieve our wishes. Thus, it is very important to express good wishes, otherwise we will observe dystopian effects. According to our view, Artificial Intelligence can be, for example, a powerful tool for improving our measure of well-being (http://www.oecdbetterlifeindex.org/), better than GDP, for analysing costs of goods and services (mainly in terms of environmental impact) and for improving and optimizing the production with the goal of minimizing, or better reaching zero impact on the Environment. “Anyone who believes in indefinite growth in anything physical, on a physically finite planet, is either mad or an economist.”— Kenneth Boulding In traditional economy, capital and work are scarce resources while natural resources have no limit.If we begin to consider that reality we live in, this is more like a spacecraft, and we have to change dramatically our vision. As an example:we have to use natural resources in a cyclic way (without waste and with respect for the Environment)we need to change our concept of growth: growth of well-being, not growth of expenses.In our vision of the world, without such a shift of perspective Artificial Intelligence may encounter problems in

							claiming ethically acceptable effects.Trying to synthesize the essence of our contribution: we suggest to integrate in the concept of AI Ethics a non traditional analysis of the socio-economic context in which AI is applied. Indeed from the awareness of such context we may better discriminated the good and the bad of AI effects and also define in a crisp way where the ethical borders are that AI researchers and practitioners should respect in their approach to AI advancements. ----------On well being see for example http://www.thefutureworldofwork.org/stories/uni-global/prioritizing-well-being-in-age-of-ai
Séverin	Tchibozo	Centre de Recherche pour la Gestion de la Biodiversité (CRGB)	No comment.	No comment.	No comment.	No comment.	It is very good initiative.
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Sean	Goltz	Global-Regulation Inc./Edith Cowan University	See comment for chapter 1 regarding clash between individuals/society/corporations/AI agents rights. Contested cases will almost always fall into this realm and therefore should be addressed. Adopting EU's treaties etc. is too generic for this purpose as they were written for humans/governments without AI agents in consideration.	"It should also be noted that, in particular situations, tensions may arise between the principles when considered from the point of view of an individual compared with the point of view of society, and vice versa. There is no set way to deal with such trade-offs." - This is a key point that should be addressed. Moreover, it is not only between the individual and society but also potential tension between a corporation and the individual etc. Most contested cases of AI ethics will fall within this realm and the			As an academic writing on AI ethics and entrepreneur using AI, I am happy to assist

				principles needs to address this situation and come up with a mechanism to resolve such challenges."AI systems that may have a subjective experience, of Artificial Moral Agents or of Unsupervised Recursively" - the document does not refer to AI agents rights and potential clash between these rights and human/society rights.			
Frans	Smit	Information Professional, supervisor, teacher and writer - (I give feedback as individual EU Citizen)	Page 2: I agree with the two components of trustworthy AI. The second component: "technically robust and reliable" is formulated too techno-centric. I think this component lacks elements like: transparency and openness of AI technology (f.e. "Well-documented": it is essential to keep the context of the AI technologies). Page 3: It is a pity that these guidelines will not be legally binding. Citizens should get more legal support to protect their rights. However this is a step forward I think. It would be great if the EU would proceed from this and constitute more regulations like the GDPR, for example concerning security and sustainability. Page 4: I like the structure of the framework, they could be implemented relatively easy into broader information governance frameworks	Page 5 and 7: I applaud the Right's Based Approach to ethics. However, I miss one fundamental right in this chapter: the fundamental right for citizens and groups to built up a trustworthy memory in the context of contemporary developments of AI. People have a right to be forgotten, but also a right to have a memory! Another right might be: ecological issues, like which type of energy to use, and to prevent pollution. This is not my expertise, so I will not elaborate on this. It should be included in the framework. Page 8: this right might for example be translated into a Principle of Sustainability: "Build up a trustworthy Memory". Page 11: Continuing on this line of resoning: I think an additional concern is Retention of code, data, description and context of AI-solutions.	Page 14: I am happy to see requirements like Accountability, Data Governance, Respect for Privacy and Transparency. I think Sustainability should be added (also regarding environmental issues). Respect for Memory should be added for example as the twin-brother of the respect for privacy. Page 15; Concerning "Design for all". I agree on this requirement af course, AI should support an inclusive society. I think the requirement should also include future generations as well. Page 17: see above for my feedback on requirement 7. Requirement 8, robustness, justly includes reproducibility. It would be good to explicitly add that reproducibility should be preserved during upgrades, migrations etc etc. It should be a platform independent requirement. Page 18: The second paragraph should include governance measures like quality management, auditing methods and PDCA (Deming) measures. I would be interested in being involved on this as well on further develop Chapter III	My feedback above might be integrated in assessment-tools like this.	I do applaud these guidelines! I think it is essential to make a stand as EU in order to enhance human centred AI, to protect and to facilitatie its citizens. I hope they will be the basis for further EU regulations on information processing. I also hope that my feedback will be a useful contribution for you to proceed to the final version.
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Regina	Hewer	none	Trust only serves as a sales argument because ethics shall not stifle … Of course, there is no legal vacuum, as there is Regulation in place, but it is neither final nor complete. There are gaps that should be higlighted and discussed. Does the "final Version of the document" imply that there is an end to discussion? Should it not be a living document as developments may come rapidly and unforeseeable? What does "sign up voluntarily" mean for the user? Will he/she use AI on his/her own responsibility? this cannot be, as the user is the weakest part in the game. Even further: How could a user take over responsibility for implementing ethics into the Technology and its use? I am wondering, what user means in this context: the human being putting ALEXA up in his Living room or Cambridge Analytica making use of the collected data. Roles and interests are definitely different ones and interests in the matter are all but the same! A more sophisticated definition is needed as there are more players in the game.	3.5 Citizens Rights does not only refer to Governments. It should also refer to companies and other private institutions. Scoring in itself is a no go as results from my right as an Individual (see also 3.2 and 5.3). All institutions be it governmental or not should inform on automatic Treatment of my data. This is also derived from my Rights as an Individual and human dignity. You should know whom you are dealing with. Do no harm: Evironmental friendly AI does not only refer to development but also to production along the whole supply chain. As long as rare metals are mined in countries like the "Democratic Republic" of Congo there is no such thing as environmental friendly electronics and also None complying with human rights. Spread of AI will aggravate the problem. Longer term concerns need close and constant watch and debate. Experience Shows that longer term could be rather short.	Data governance: it should not only be ensured that the data are not used against their providers - Data should only be gathered if the Providers gave their explicit consent to gathering them.		
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential

Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Mihai	Voicu	EU citizen	S1. (page ii) The Guidelines (GL) should have a well-balanced individual vs community approach and it should stress prosperity of mankind.S2. (page ii) Chapter II should include provision on non-decreasing economic and social power of masses. (see General Comment G5.)S3. (page iv, AI definition) Please assess following AI comprehensive definition elaborated from different sources: "humanised portable or not virtual assistant application or humanised or not virtual assistant machine shape platform that performs one or multiple intellectual task(s) that a natural person or animal can, including but not limited to, reasoning, solving problems, decision-making, knowledge, planning, learning, recognizing, natural language processing, speaking, gesturing, mimicking, perception, rationally move (from place to place) and move and manipulate objects".S4. (page iv, Trustworthy AI definition) There is a need to clarify to which fundamental rights Trustworthy AI refers to. (see Specific Comment S5.)S5. (page 1, para 1 and 3) European values and fundamental rights should be clearly stressed by relevant footnotes.S6. (page 2, para 1 and 2) Future relevant AI regulatory framework (directives, regulations, recommendations, opinions, rules, laws, norms etc.) should tackle the whole lifetime cycle of AI.S7. (page 2, The Role of AI Ethics section) There is a need to clarify the goal of AI ethics from the 3rd sentence of the first para of The Role of AI Ethics section.S8. (page 3, Scope of the Guidelines) GL should refer to AI as a whole not to AI apps or AI Systems, which are not defined in GL.	S9. (page 7, para 3.1) GL should refer to AI as a whole, not to AI apps or AI Systems, which are not defined in GL.S10. (page 7, para 3.1) AI should treat humans with respect instantly and forever.S11. (page 7, para 3.2) AI should protect economic and social power of masses.S12. (page 12, The Principle of Non-maleficence) AI should protect at all costs the dignity, integrity ... and security as well as health, economic and social power of masses and planet and planet climate.	S13. Further AI usage in economic, social, educational and health processes will limit human participation as labor force to these processes or banish humans from these processes (as AI is more efficient and effective) and masses will lose their current economic and social power as their power will concentrate in hands of AI developers/owners. So, GL should tackle (further) these aspects. S14. (page 14, Requirements for Trustworthy AI) Requirements for AI user reachability, scalability and interoperability should be added.S15. Requirements for AI to protect always health, economic and social power of masses should be added.S16. Requirements for AI to protect always planet and planet climate should be added.S17. Requirements for AI developers/owners for disclosing self-assessment against GL should be added.	Please assess the following requirements:Access policies 1. When providing an AI technology/product/service, an AI owner should ensure that its AI technology/product/service is reachable by potential AI users, namely consumers and firms.2. When providing an AI technology/product/service, an AI owner should ensure adequate scalability by its AI technology/product/service, taking into account AI and innovation developments.3. When providing an AI technology/product/service, an AI owner should ensure adequate and complete control of its AI technology/product/service, even when outsourcing or providing the AI technology/product/service in cooperation with other entities.Risk management4. When providing an AI technology/product/service, an AI owner should comply with relevant legal framework, including relevant competent authorities' recommendations and warnings, taking into consideration different jurisdictions.5. When providing an AI technology/product/service, an AI owner should aim to comply with relevant high standards and best practices.6. An AI owner should closely monitor AI developments and risks posed by these developments to its business, on regular basis.7. An AI owner should develop and promote a responsible AI innovation culture.8. An AI owner should define, identify and support responsible AI innovation (which allows and promotes EU human values and rights, health, education, financial stability, adequate comprehensive and complex risk management, competition, etc.).9. An AI owner should include responsible AI and innovation in its development strategy.10. An AI owner should include AI and innovation vulnerabilities and risks within its risk management framework.11. An AI owner should review risk management framework concerning AI developments (especially towards operational and cyber risks, including silent cyber risk) on regular basis, in order to maintain a sound risk management framework, even when outsourcing.12. An AI owner should regularly review risk management framework based on lessons learn from impact of AI and AI innovation developments to own business and third parties.13. When providing an AI technology/product/service, an AI owner should allot appropriate resources.14. An AI owner should regularly review its resources in order to ensure that all resources remain available and fit to the complexity and evolution of risks and vulnerabilities posed by its AI technology/product/service, including but not limited to (1) experienced skilled managers and staff, (2) appropriate and enforceable policies and procedures in diverse and extreme (but plausible) scenarios, (3) assets and (4) appropriate and enforceable business continuity arrangements.15. An AI owner should regularly review its knowledge base in order to ensure it has always a deep knowledge of the AI technology/product/service.16. When providing an AI technology/product/service, an AI owner should assess AI impact on its risk appetite and to third parties, taking into consideration possible inadequate risk	General CommentsG1. These Guidelines (GL) should end up as a support for EU regulatory framework for setting up rules for both AI usage and interaction between humans (natural persons) and individual AI (future conscious or not artificial persons). G2. GL should have a well-balanced individual vs community approach and it should stress prosperity of mankind.G3. AI definition needs further regular improvement (see specific comments) and there is a need to use only one term when referring AI, instead of AI, AI systems or AI applications.G4. There some trials where AI platform(s) is/are used for developing systems, so it is possible that AI will be involved in developing AI.G5. Further AI usage in economic, social, educational and health processes will limit human participation as labor force to these processes or banish humans from these processes (as AI is more efficient and effective) and masses will lose their current economic and social power as their power will concentrate in hands of AI developers/owners. So, GL should tackle these aspects. G6. GL should stress AI impact to health, education and social (including debating and voting) and GL should stress further these issues. GL should stress AI impact to climate and planet and GL should stress AI climate sustainability.G7. There is a need to clarify the goal of AI ethics from the 3rd sentence of the The Role of AI Ethics section.G8. Further areas of Requirements of Trustworthy AI should be tackle including but not limited to: AI interoperability, scalability and reachability to potential AI users (latter differs from "design for all" concept – which is comprehensive).G.9 GL should contain provisions regarding the whole lifetime cycle of AI.G10. GL should contain provisions regarding AI developers/owners disclosing results from self-assessment against GL.G11. GL should be revised on regular basis.

assessment.17. When providing an AI technology/product/service, an AI owner should ensure both AI user interests protection (including AI consumer long-term health/life safety) and environment protection.18. When providing an AI technology/product/service, an AI owner should ensure adequate balance between strong security and user convenience (user-friendly environment), taking into account AI and innovation developments.19. When providing an AI technology/product/service, an AI owner should ensure appropriate incident and crime detection and prompt disclosure relevant authorities.20. When providing an AI technology/product/service, an AI owner should ensure always a high resilient level of its AI technology/product/service, including adequate measures to protect locations, development/testing/production environments and networks, websites, and communications against abuse or attacks, taking into account (1) diverse and extreme (but plausible) scenarios, (2) evolution of vulnerabilities and threats and (3) counter-measures developments.21. When providing an AI technology/product/service, an AI owner should ensure regular and adequate testing of its AI technology/product/service, taking into account (1) diverse and extreme (but plausible) scenarios, (2) evolution of vulnerabilities and threats (3) counter-measures developments and (4) AI and innovation developments.22. When providing an AI technology/product/service, an AI owner should ensure adequate efficiency and sustainability by its AI technology/product/service, taking into account AI and innovation developments.23. When providing an AI technology/product/service, an AI owner should ensure its AI technology/product/service does not take advantage of humans (e.g. induce human hardship or enslave).Interoperability24. When providing an AI technology/product/service, an AI owner should ensure adequate interoperability of its AI technology/product/service (including formalised AI testing programs and environments), taking into account AI and innovation developments.Communication and Education25. When providing an AI technology/product/service, an AI owner should ensure regular proactive appropriate updated and fair AI user information (awareness) and education, including AI technology/product/service usage, risks and treats posed by its AI technology/product/service and name of its IT/cyber-security experienced auditor(s).26. When providing an AI technology/product/service, an AI owner should ensure prompt disclosure of incidents and crimes related to its AI technology/product/service to its AI users.27. An AI owner should ensure that communication with AI user is based on clear non-misleading easily readable formats in natural daylight environment.28. An AI owner should ensure at least one end-to-end secure communication environment dedicated to AI user.29. An AI owner should disclose the meanings of relevant technological concepts, by means of a glossary, but taking into consideration confidentiality and intellectual property issues.Cooperation30. An AI owner should promote and support cooperation with both

						private sector and public sector (relevant domestic, regional and international authorities) where possible.31. An AI owner should publicly disclose compliance to all these requirements, by means of self-assessment compliance report.	
Anonymous	Anonymous	Anonymous	We need good rules for the humans, not the companies!			I think in a Long it is difficult to rule an AI. An AI might become in certain Areas more intelligent then humans. How do we keep the lead?	I demand for all AI who assess me (e.g. credit rating) a full insight of the arithmetic and the the input data.
Benedikt	Blomeyer	Allied for Startups	Response to Consultation: Draft Ethics Guidelines for Trustworthy AIAllied for Startups welcomes the draft ethics guidelines from the High-Level Expert Group on Artificial Intelligence, in particular its positivity and future-oriented character, and would like to contribute to an innovation and entrepreneurship-focused debate.	Regarding conceptual clarity, being treated respectfully as an individual and being a data subject are not mutually exclusive (Chapter I. 3.1: Respect for Human Dignity).	We recommend prioritising clarity with terminology and concepts, to the extent that the guidelines are understandable for an entrepreneur. It will be challenging to distinguish between personalisation and 'extreme' personalisation, or to understand what constitutes 'individual choice' (Chapter II. 1.6: Respect & Enhancement of Human Autonomy). Another example is the reference to 'human data' or a 'morally significant impact' (Chapter II. 1.10: Transparency). Legally ambivalent concepts will lead to costs and complexities that are not negligible for entrepreneurs. The guidelines should not be a legally nuanced text for experts, but intelligible for entrepreneurs.		The sheer breadth and depth that the development and application of Artificial Intelligence offers can seem overwhelming. It invites taking a step back and asking more fundamental questions, such as: Why is AI being developed; Who are the practitioners who drive AI and dare to moonshot these ideas? In this process, doomsday scenarios can seem to address real fears people may have, blaming technology for unrelated problems. It is one reason why AFS welcomes the distinctly positive approach that these draft guidelines have taken. They form an invitation to a constructive dialogue, which we believe is necessary at such a formative time for AI. As the guidelines rightly stipulate, it is neither desirable nor possible to provide a precise AI cookbook. Instead of prescriptive or technical instructions, a principled approach is chosen. It allows cost-benefit analysis on a case by case basis, leading to a tailored approach.A principled approach is desirable and it should be measured against startups' abilities to enter a market and compete in it. Allied for Startups has long argued that AI needs to be understood through startups. As the smallest, innovative entities, they are the

							one's thinking about new opportunities and business models all the time. There is no "AI made in Europe" without startups. If these guidelines lead to highly bureaucratic and front loaded obligations, many entrepreneurs will think twice about their next startup. In other words, getting trustworthy AI by design is best achieved by making guidelines that inspire entrepreneurs to take bold decisions and think the unthinkable - with the human at the centre. Startups are global from day one. A too strict definition along the lines of a 'made in' label contradicts the global character of startups. Many of them could be inspired abroad, try to refine their business model by learning from others, or build on something tried elsewhere. Oftentimes the best products and services aren't produced in one country, but are based on a series of learnings and components from across the world. Instead, recognising that AI develops and grows globally means that there needs to be a global conversation. In areas of synergies, such a conversation can lead to a strong global community, in others it can help to identify European excellence. At the end of the day, such a conversation can also lead to a discussion on global governance of AI.In closing, we encourage experts to consider the comprehensive corpus of European laws that these guidelines will be complementing. When a new AI application emerges, it might not need not need a new law, but maybe an overhaul or an application of existing laws. In that light, we urge keeping laws simple, evidence driven and specific.
Maria Luisa	Guerrini	Ordine Degli Architetti Della Provincia Di Perugia (Order Of Architects Of The Province Of Perugia)	AI must help us to EXPAND CREATIVITY, to go beyond today's human limits, but certainly not to REPLACE IN CREATIVITY. Is necessary to ensure that the AI can not access the rights of Copyright or other legal institutions that sanction their creative autonomy: they would replace us and not strengthen us, trampling in this way our DIGNITY (element that the guidelines pose as primary factor to defend).	3.1 – Respect for human dignity Creativeness is one of the major "intrinsic worth" possessed by human being. An AI respect human dignity if serves creativity of artists or professionals, not if it stands in for them. An AI cannot be creative, in legal and cultural sense of identity. Creativeness is an exclusive human being value. 5.2 – Covert AI system The humans being (citizens or consumers) must be aware, when they are interacting, buying or enjoying of activities creativeness, if they were produced by artists/professional or by AI. The confusion between human and AI creativeness has multiple consequences such as the reduction of the intrinsic value of human being, in particular.	2 – Non-technical methods – Regulation In compliance with fundamental rights (Chapter I: creativeness), the activities or goods or services based on creativity stemming from an AI cannot be protected by copyrights or others kind of patents. – Education and awareness to foster an ethical mind-set In the category of the "users" (companies or individuals) must be included the councils or the organizations of artist and professionals that working with the creativeness. – Stakeholder and social dialogue In the category of the "stakeholder" must be included the councils or the organizations of artist and professionals that working with the creativeness.	Check points aimed to the protection of human creativeness is absent in: 1. Accountability 2. Data governance 7. Respect for human autonomy 8. Robustness – (Accuracy through data usage and control) 10. Transparency – Purpose	We consider a priority to protect the CREATIVITY of those who carry out intellectual work, so we submit our contribution.

Marc	Steen	TNO

First, I would like to congratulate the AI HLEG with this document. It's clear, it's accessible, it's thorough, and it's practical. Let me sum up all the things I find brilliant:

They use 'Trustworthy' as an overarching term. I think this is brilliant. No matter how you conceptualize AI–as 'general AI' or 'narrow AI', as 'AI in autonomous systems' or 'AI as a tool to advance agency of humans'–we can all relate to the need for AI that is worthy of our trust. You want a trustworthy AI similar to how you want a trustworthy car, a trustworthy drilling machine, a trustworthy babysitter, or a trustworthy partner.

They explain the relationships between rights, principles, and values. Rights provide the "bedrock" for formulating ethical principles. And in order to uphold these principles, we need values. Moreover, we need to translate rights, principles, and values into requirements for developing AI systems. Putting rights, principles, and values into these relationships provides clarity, which is direly needed for a constructive discussion of ethics. They discuss the following rights, principles, values, and requirements:

They structure their guidance in three parts, from abstract to practical: Guidance for ensuring ethical purpose; Guidance for realizing trustworthy AI; and Guidance for assessing trustworthy AI. Such a structure is very useful, and much needed, during the design process (purpose), implementation process (realizing) and evaluation process (assessing). We need to move from abstract to practical, and back, in an iterative fashion—indeed, in iterative cycles.

Concern for Human Dignity

The AI HLEG asked for feedback on "Critical concerns raised by AI" (pp. 10–13). I would like to propose to add one concern: concern for human dignity.

What do I mean by that? Well, you are familiar with the Turing Test. It aims to evaluate whether a computer can give a performance that we recognize as human-like intelligence so that we cannot distinguish it from a human. In a Turing Test the computer's aim is to behave like an intelligent person.

Now imagine a Reverse Turing Test. In such a test you, as a human being, aim to adapt to the computer and its algorithms. You fix your eyes on your mobile phone's screen and you mindlessly click 'okay', 'view next', 'buy'–you do whatever the algorithm tells you to do. In a Reverse Turing Test, your aim is to behave like a machine.

This concern is related to other concerns discussed by the AI HLEG: for 'Identification without Consent' (when you mindlessly click 'yes, I accept terms and conditions'), for 'covert AI systems' (when a system treats you in a mechanical manner, with machine logic), and for 'Normative and Mass Citizen Scoring' (when a system gathers all sorts of personal data and uses these for all sorts of purposes, in non-transparent ways).

Implementing too many AI systems, in too many spheres of life, and using these too much, is a threat to human dignity.

This concern was discussed, e.g., by Brett Frischmann and Evan Selinger (Re-engineering Humanity, 2018: 175–183; I took the idea for a Reverse Turing Test from them), by Sherry Turkle, who reminded us of the value of genuine human contact, both intra-personal and interpersonal (Reclaiming Conversation, 2015), and by John Havens (Heartificial Intelligence, 2016), who advocated "embracing our humanity to maximize machines": to design and use machines in ways that preserve and support human dignity.

Putting Human Agency First

Furthermore, I'd like to propose an improvement and clarification in the formulation of two of the 'Requirements of Trustworthy AI' (pp. 13–18).
The AI HLEG discusses "Governance of AI Autonomy (Human oversight)" and "Respect for (& Enhancement of) Human Autonomy". My proposal is to merge these requirements into one requirement, under the heading of, e.g., "Appropriate Allocation of Agency", or: "Putting Human Agency First".

Both requirements ("Governance of AI Autonomy" and "Respect for Human Autonomy") are about distributing agency between people and an AI system. Put simplistically:

-Moral agency resides in people, not in machines;
-there are only 100 agency-percent-points to share (as it were);
-and you can delegate some agency-points to a machine;
-but then you will lose these (like in a zero-sum game).

The agency of humans and the agency of an AI system are on one and the same axis: on one side of this axis people have 90% of the autonomy and the AI system 10%; on the other side the AI system has 90% of the autonomy and people 10%. The choice is ours—and we will need to decide carefully, taking into account the various pros and cons of delegating agency to machines.

Merging these two requirements about autonomy is intended to clarify that human agency diminishes when we delegate agency to machines.

Underlying this intention is the belief that technology must never replace people or corrode human dignity. Rather, we need to put human agency first, and use technologies as tools. Here it needs to be acknowledged that tools are never neutral; the usage of any tool shapes the human experience and indeed the human condition (https://ppverbeek.wordpress.com/mediation-theory/)—this requires careful decision making, e.g., in the ways in which an AI-tool gathers data, presents or visualizes conclusions, provides suggestions, etc.

This idea is at the heart of the Capability Approach, which views technologies as tools to extend human capabilities: to create a just society in which people can flourish (http://www.mitpressjournals.org/doi/abs/10.1162/DESI_a_00412).
This idea is also expressed in the "Statement on Artificial Intelligence, Robotics, and 'Autonomous' Systems" of the European Group on Ethics in Science and New Technologies, in which 'Autonomous' has quotation marks to indicate that a system cannot have moral autonomy. Finally, the principle of "appropriate allocation of function between users and technology" is explicitly mentioned as a principle in the ISO 13407:1999 standard for Human-centred design processes for interactive systems (the updated ISO 9241–210:2010 standard puts this less explicitly).

Virtue Ethics for Human Flourishing

Moreover, the AI HLEG invites suggestions for technical or non-technical methods to achieve and assess Trustworthy AI. In line with the suggestions above (a concern for human dignity; and putting human agency first), I'd like to propose to add virtue ethics to the mix of non-technical methods.

In her book "Technology and the Virtues" (2016), Shannon Vallor advocated developing and using technologies in ways that promote human flourishing. She views technologies as tools that can help—or hinder—people to cultivate specific virtues. She argues that we need to cultivate specific technomoral virtues to guide the development and the usage of technologies, so that we can create societies in which people can flourish in the 21st century.

Please note that each society, for each specific era and area, needs to make its own list of virtues that are needed for that society. The virtues that Aristotle proposed were for the citizens of ancient Athens. The virtues of Thomas of Aquinas were for medieval catholic people. Vallor proposed the following virtues for our current global, technosocial context (op.cit.: 118–155):

Honesty (Respecting Truth), Self-control (Becoming the Author of Our Desires), Humility (Knowing What We Do Not Know), Justice (Upholding Rightness), Courage (Intelligent Fear and Hope), Empathy (Compassionate Concern for Others), Care (Loving Service to Others), Civility (Making Common Cause), Flexibility (Skillful Adaptation to Change), Perspective (Holding on to the Moral Whole), and Magnanimity (Moral Leadership and Nobility of Spirit).

Vallor argued that virtue ethics is an especially useful approach for discussing the development and usage of emerging technologies (op.cit.: 17–34): technologies that are under development and not yet crystallized. AI is an example of an emerging technology. Emerging technologies entail what Vallor calls "technosocial opacity" (op.cit.: 1–13); their usage, integration into practices, effects on stakeholders, and place in society are not yet clear. She argues that other well-known ethical traditions, like deontology or consequentialism, can have limitations when used for the development and usage of emerging technologies. In deontology, one aims to find general rules and duties that are universally applicable. In consequentialism, one aims to maximize positive effects and minimize negative effects for all stakeholders. For an emerging technology like AI, however, it is hard to find general rules and duties, or to calculate all possible effects for all stakeholders (op.cit.: 7–8).

Take, for example, autonomous cars—with lots of AI in them, and in the infrastructure around the cars. Yes, there are some cars driving around with some level of autonomy. But they are not fully autonomous and they are not widely used. Therefore we cannot yet have a good-enough understanding of the ways in which people use autonomous cars and of their place in society.

Autonomous cars may, e.g., incentivize

I submitted a very similar response earlier in December. I also posted it on Medium (https://medium.com/@marc.steen/ethics-guidelines-for-trustworthy-ai-to-promote-human-dignity-agency-and-flourishing-a664f000c5a5), where it sparked some discussion. This helped me to improve my response. The current version can replace the earlier one.

							people to make longer commutes: to travel 4 hours in the early morning (while sleeping behind the wheel) and travel 4 hours in the evening (while watching videos). This could disrupt family lives, corrode leisure time, social interactions and the social fabric of society, and have huge negative impacts on the environment—and on traffic congestion. For such a case, it would be hard to know exactly which duties are involved or which general rules apply. Or it would be hard to anticipate and calculate all the positive and negative consequences for all stakeholders involved. A virtue ethics approach, however, would be useful here: to identify the virtues that are relevant in this specific case (to create a society in which people can flourish), and to provide recommendations to cultivate these virtues, including processes of self-examination and self-direction (op.cit.: 61–117). Rather than putting different approaches in opposition to each other, to disqualify one, or to favour one at the expense of another, I'd like to propose to create a productive combination: to use deontology where and when we have clarity about general rules and duties; to use consequentialism where and when we are able to calculate positive and negative consequences; to use virtue ethics where we ask questions about what kind of society we want to create and how technology can support people's flourishing.	
David	Pereira	everis, an NTT Data company						The report defines 10 requirements of Trustworthy AI : 1. Accountability 2.Data Governance 3. Design for all 4. Governance of AI Autonomy (Human oversight) 5. Non-Discrimination 6. Respect for (& Enhancement of) Human Autonomy 7. Respect for Privacy 8. Robustness 9. Safety 10. Transparency. We would add to the list "Fair competition" (Taken from the Japanese government's AI 7 principles): • Certain countries or enterprises should not monopolize data and/or concentrate most of the wealth generated by AI.
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential

Thomas	Bolander	Association of Nordic Engineers	Purpose and Target Audience of the Guidelines ANE welcomes the idea of a concrete mechanism enabling stakeholders to formally endorse the Guidelines, which is a good incentive for a non-binding commitment towards the trustworthy AI. In ANE we consider that Europe should take the step further and create an ethical certification for AI systems. Ethics should go hand in hand with the brand of the establishment and thereby be an integral part of codes of conduct of each institution working on an AI system. The EU "AI-friendly label" attribution could be a reward for AI trustworthy organizations and a good trademark for Europe leading the way on AI and ethics.	5. Critical concerns raised by AI ANE draws attention to the fact that the list of critical concerns raised by AI, page 11, could be supplemented with the political race of artificial intelligence and use of data-mining tools, which are used to influence political decision making, exercise control, infringe on the freedom of expression and the right to receive and impart information without interference. Ethical considerations might not be the highest priority for the world leaders competing on AI and using data-mining tools to undermine each other's credibility and privacy.	While recognising the exhaustive list of requirements, page 14, ANE wishes to add an additional element to the requirement on accountability. There should be developed a clear appeal process and instance with governmental oversight. Such a process must enable individuals and organisations to address the AI behaviour and decisions that they find potentially harmful. Additionally, the list of requirements should include the notion of trust. Gaining trust and ensuring accountability goes together, and both questions are deeply connected to the structures of the organisations that produce particular technologies. There is no recognition in the document that trustworthiness in part stems from the socio-technical institutional structures within which these systems are deployed. Hence, it would be wrong to assume that trustworthiness only equates to particular features of the systems themselves. When talking about transparency, it is important to underline its limitations, which are merely touched upon in the current draft. Efforts towards transparency can often produce so much information that what is important can be made obscure in the deluge unintentionally or attempts at transparency do not necessarily result in building trust. While transparency is a worthwhile goal, its application requires considerations of potential pitfalls. According to ANE three components are worthy of consideration: 1) embedding "Transparency by Design" throughout the development process; 2) establishing an independent verification body and 3) providing transparency of the decision-making processes at the organisations responsible for building the AI technologies. Moreover, ANE considers that the current list of requirements should include an additional requirement, namely addressing bias. Given the attention currently being paid to the importance of training data used in the development of any AI system that relies on algorithmic data processing, it is crucial to ensure that these considerations are addressed in practice. As bias is hard or impossible to remove, the emphasis should be put on awareness (revealing) and auditing biases. This could be done by: 1) establishing internal training programmes for staff to deepen skills for ethical reflection and recognition of biases, and 2) putting in place an external, neutral entity, such as "a model testing institute" to repeatedly audit AI systems. ANE supports fully the idea of promoting education and awareness to foster an ethical mind-set, page 22, and suggests broadening the current proposal. Efforts to augment or even reform technical education is already happening at different levels, but this development is happening either through grass-roots efforts or with the support of civil society and commercial actors. For these changes to become systematic however, it is clear that government support and oversight are crucial. Furthermore, ANE is pleased to see the reference to social partners in the draft proposal. Yet, all stakeholders need spaces for sustaining a living dialogue around issues	

of AI and ethics. Supporting such deliberations and dialogue must not fall exclusively on the shoulders of the relevant stakeholders themselves but requires sustained political backing and government investment to be sustained. There is a need to create formal and informal structures enabling an open, public and ongoing debate across and between stakeholders on the development, deployment and use of AI systems.

Concerning the technical methods, pages 19-21, ANE emphasizes that efforts towards AI development must recognize the socio-technical nature of the development process. Expanding the disciplinary orientation of AI research will ensure deeper attention to social contexts, and more focus on potential hazards when these systems are applied to human populations.

On page 20, the current text refers to: "… a stochastic system is often described by a "sense-plan-act" cycle. For such architecture to be adapted to ensure Trustworthy AI, ethical goals and requirements should be integrated at "sense"- level ..." Sensing means getting input from the environment, so isn't it rather at the plan-level that ethical goals and requirements should be implemented? Plans should be formed to adhere to certain ethical requirements (constraints on allowable plans).

Regarding testing & validating on page 20, ANE recommends the text to include considerations of formal verification. Many modern IT systems, including microprocessors, e-voting systems and train signaling systems, are formally verified. A mathematical proof is given that they are flawless and have the required properties. Formal verification can also be used on AI-systems. Machine learning systems themselves cannot be formally verified, but they can be encapsulated in other systems that process the output of the machine learning system and for instance only allow certain action to be executed. Testing and verification are not the same, but for safety critical systems, verification is often needed or at least thrived for.

The explanation (XAI research) on page 21 would largely benefit from the better integration of symbolic and subsymbolic AI techniques (e.g. neural nets). Contrary to learning systems based on neural nets, symbolic AI systems like rule-based systems or planning systems tend to be very well fit to provide clear reasons for their decisions. However, these of course lack the learning aspect. As with humans, AI systems need both to be able to perceive and categorise sensor input (neural networks), but also to take complex decisions and use a language to explain those decisions (symbolic AI).

Marc	Le Goc	Aix-Marseilles University

Page 5, §2, about the notion of "informed consent": Informed consent according to a Principle of rationality being a difficult tasks for anyone with usual information systems, it is much more difficult, if not impossible, with AI systems. This point would be deeply questioned!Page 7, §3.2, about the protection of the freedom of individual persons: how to verify that the value added by AI technologies will be equitably redistributed? It seems to be some thing impossible!Page 11, §5.1, about the identification of individuals: AI systems can (and must) work on numbers only, without any reference to the individuals behind these numbers : the link numbers-humans MUST be done by human beings, and only human beings, under the command of the legal responsible of the organization.

Page 14, §1, about accountability: Up to my opinion, this is the most important point about the requirements of Trustworthy AI. Remember the attitude of Mark Zuckerberg face to the different democratic commissions in USA or Europe : he is accountable of nothing ... because he know nothing about Facebook processes! It is a sad joke but ... it works! So the problem lies not in compensation mechanisms but in forcing the organization's leaders to assume the responsibility of the harmful consequences of the AI systems. Knowledge is power, and with power comes responsibilities!Page 17, §8 about Robustness: Robustness is one of the most technical difficulties with AI systems because it is the condition for reliability and reproducibility. This is not a question of complexity nor opacity: any AI-system is deterministic since it runs in a Turing Machine, except when a kind of random function is included in the knowledge model. But, I never seen an operational AI decision system using a random function: only off line learning systems uses random functions to avoid local extrema and so achieve reproducibility (not reliability which is another problem depending on the representativity of the learning and testing data). When the parameter's model have been fitted, the behavior of any AI system must be determinist because if not, no operational can be reached!Page 19, §1 about Ethics and Rule of law by design: The key point here is the traceability of the knowledge pieces from its formulation in natural language to the final pieces of code and parameters. This is easy to do and should be an obligation for any AI system ... but Artificial Neural Networks failed on this point!Page 20, §1 about Testing and validating: Here again, the key point is the traceability of the knowledge pieces from its formulation in natural language to the final pieces of code and parameters ...Page 20, §1 about Traceability and audibility: The notion of causality is quite strange when considering a computer system. May the term of logic should be more appropriate here because most of the AI system uses non-causal algorithms, only real time systems use causal algorithms. And classification system like Deep Learning systems are not causal: they are simple associative memories, that's all!

Page 24, §III about the primary target audience of the Assessing trustworthy AI chapter: The responsibility of Designers, Team Leaders or Developers are limited by their employer, their boss, their enterprise or organizations: they can’t go against them! Only the manager must be personally accountable for AI systems. This chapter should recall the obligations and the rights of Alert Launchers which is the only way to avoid excessive ambitions of managers. Knowledge is power, and with power must come responsibilities!Page 24, $1 about Accountability: Who is accountable if things go wrong? The top manager of the organization of course! No doubt on this. And again, Alert Launchers are of the most importance to avoid the things to go wrong!Page 25, §4 about Governing AI autonomy: the question 4 about the measures to be taken to ensure the AI system decisions must be completed with the affirmation that the proof of an error is always the charge of the system owner.Page 25, §6 about the respect for privacy: The question "If applicable, is the system GDPR compliant?" is chocking! By definition, the GDPR is compliant! I don't understand why such a question is made in a document about ethics in a software system!Page 26, §7 about the respect for human autonomy, question 4: It is clearly and definitively impossible for a user to interrogate the algorithmic decision: real AI systems are, by definition, too much complex software's for that! Even with usual information systems this is not possible! Here again, the proof of the respect for human autonomy is the only charge of the system owner, never the user!Page 26, §8 about Reliability and reproducibility: The first question is very strange : if the system does not meet its goals; purposes and intended applications, its is not operational and can’t (must not) be used!!!!!Page 26, §8 about Reliability and reproducibility: The term "mechanisms" of the last question is dangerous. It is easy for anyone to introduces mechanisms some where to defend the fact that all has been done to assure users of the reliability of an AI system! The question is not the mechanisms but the proof of their efficiency!Page 26, §8 about Accuracy: As evoked in the introduction, the problem is not accuracy : the problem is certainty.Page 26, §8 about Accuracy, question 3: Looking for the completeness of data is a non-sens. Only reprentativity is relevant.Page 26, §8 about Accuracy, question 4 and 5: This is the first time that notion of model is used in the document. Sorry but this show a strange understanding of what AI systems are and the crucial role of knowledge models in AI.Page 27, §10 about the purpose: The question of who may benefit from AI systems is crucial and fundamental. Refer to the Homo Deus notion of Yval Harari to understand the importance of this point for democracies.Page 27, §10 about the purpose: The limitations of AI system will never be specified to users and there is a good reason for that : do you know the limitations of the Deep Learning AI systems? Officially, they have no limitations ! Nevertheless, here are the 3 mains limitation for face recognition systems: a zoom less than 10, a rotation less that 3°c and less than 1% of noise. if the public knew these limits, they would never trust this kind of system, of course! Page 27, §10 about

The first point to recall is that a AI system being a particular (restricted) sub-set of the set of all the software's, it always belongs to some body (a company, an association, a government, a startup, a person, etc). A software, and then any AI system, always has an owner. As a consequence, the responsibility, at the end, must always lie on the owner of the AI system: this is the only way to enforce the application of theses Ethics Guidelines.The second point is that an AI system is a special kind of software because it manipulates a strange and ill-defined matter: knowledge. Recall Newell's definition of knowledge (1982): knowledge is all that can be imputed to an agent, so that its behavior can be assessed against the principle of rationality. This means that if a system acts according to a principle of rationality, then it is legitimate to say that it uses knowledge. And Newell to precise: "Knowledge must be functionally characterized, in terms of what it does (its role), and not structurally, in terms of physical objects with particular properties and relationships". Since Newell, any Knowledge Engineer knows that the most important aspect of knowledge is the role it plays in a problem solving method.The third point is that the notion of Knowledge requires to refer to an interpretation, a human being interpretation in this context. We know since Shannon (1948) that Data in not Information, and Information is not Knowledge. Data is raw (or physical) material so that a "Data Base" contains no Data but Information only. Information is coded Data, a representation of Data according to a coding system always based on the Bool set B={0, 1} in this context. As a consequence, Information is an abstract representation of Data. Knowledge is more complicated to define because this notion adds a subjective point of view on Information. My personal definition is the following: "Knowledge results from an intentional interpretation of a flow of information". This definition is inspired from Damasio's works. And again, Knowledge is more abstract than Information. So, according to Floridi's Method of Levels of Abstraction, Data, Information and Knowledge are located on three different levels of abstraction, themselves organized in a nested gradient of abstraction. It is then a pity that all along the text, the confusion between Data, Information and Knowledge is constant, permanent: in the necessity of the interpretation lies the all the ethical biases.Theses recalls are sufficient to argue that AI system are characterized by the fact that, as a cognitive agent, it uses knowledge to reach its goals. The key point about AI system comes from the fact that "knowledge is power " (not information nor data but knowledge only). And the power does not give itself, it takes itself! That is to say that access to power by the mean of knowledge can lead to war, and usually do in the History. To avoid power's wars must be the first order of these ethics guidelines.This leads to my main concerns about this text: Data is not the main feature with AI system, but Knowledge and Knowlegde Models are. And in this text, may be because of the recent popularization of Deep Learning technology, the emphasis has been put on data not on knowledge. This is the obvious biases in the analysis of the AIHLEG: information is not knowledge, information is

traceability: The question 1 (about the measures to inform on the accuracy) has nothing to do with transparency but accuracy.Page 27, §10 about traceability: Concerning the question 3 and the method of building the algorithmic system. There is something confusing with the term « algorithmic system » when talking about AI systems : all problem having no algorithmic solution requires an AI approach! More, why making a difference between rule-based AI systems and learning-based AI systems? After all, there existes current researches that provides learning algorithms able to automatically build the set of rules directly from data, and the learning algorithm used to build a learning-based AI system produces also a knowledge model. Even my students know that! Strange misconception of what an AI system is ...

a necessary condition for knowledge creation but definitively not a sufficient one. Knowledge is power and with power comes responsibilities. As a consequence, up to my opinion, the only way to achieve Trustworthy AI and human centric AI systems is to affirm the two indisputable following points.The first indisputable point is that an intelligent machine will never be responsible for its acts: only the owner of the control software of an intelligent machine is responsible for the eventual damages. Recall Elaine Herzberg, killed with an autonomous Uber car in 2018.The second indisputable point is to force to make explicit the Knowledge Models used in any AI system (the inference engine and the learning algorithms are generally well documented, so they are usually much more clear that the embedded knowledge models). This is where the main focus must be put on: the knowledge models, the only media allowing ontological and epistemological assessments.And this is where fail Artificial Neural Networks (ANN or Deep Learning if you prefer): up to now, there is no way to explicit the knowledge models built by ANN learning algorithm! Up to my humble opinion, this intrinsic limitation of ANN or Deep Learning system explains the biases in the text of AIHLEG.But it is not because no solution exists till today (even this assertion is false!) that the focus must be mainly put on "data" (learning or testing data bases, which contains information, remember). Clearly, knowledge models are definitivly required to build the data bases: generally speaking, these models are oblivious, unconscious, but the are required to build a database.So all the questions of the AIHLEG about "data" would (should) be formulated about knowledge or knowledge models used to to design any AI system, with or without learning or testing data bases.The usual, classical and old argument against such an elicitation is the cost (the famous bottleneck of knowledge modeling): building a knowledge model is always an expensive operation. But if this argument was true in the past, it is not to days: a lot of progress have been made during the last decades to describe the contain of a data base, notably in the Data Mining or KDD (Knowledge Discovery In Database) domains. There is then no more legitimate reason to avoid to require to the owner of an AI System to explicit their knowledge models.If not, is it responsible to accord our confidence in a pure black box?What would effectively means a Trustworthy AI in that case?Would you confide your children to a bus driver who is a notorious mental patient?My answer is no: and the only way to assure that an AI system is not crazy is the elicitation of its knowledge models.I know that Deep Learning make the buzz since the works of Hinton on Artificial Neural Networks in 1999. But it is not the only researcher in AI having made significant contributions since these last 40 years: Deep Learning make the buzz thanks to the "GAFA". But Deep Learning is only one may to build AI systems.Only one example.Let me cite the French cognitivist, neuroscientist and psychologist Stanislas Dehaene (http://www.college-de-france.fr/site/stanislas-dehaene/) about the simplicity and the efficiency of Bayesian Inference (and then Bayesian Networks AI systems widely used in autonomous vehicle for example) : "Bayesian Inference is one of

the main reasons for the "Bayesian revolution in Cognitive Sciences" where it is widely used to model a very large diversity of cognitive phenomena: perception, statistical inference, decision-making, learning, language processing, ...".Deep Learning based AI systems don't uses Bayesian Inference. It implements one and only one cognitive operation: classification. According to Stanislas Dehaene, what is a classification system that don't use Bayesian Inference? It can be a human-like classification!So, what are Deep Learning based AI systems? Are they true AI systems? Sorry for the defenders of Deep Learning but the question is legitimate. And according to cognitivist's, neuroscientist's or psychologist's, the answer is definitively no.Strange situation, isn't it?Clearly, my intention is not to argument against Deep Learning based true AI systems. My intention is to justify my opinion about the constant, permanent and obvious biases all along the AIHLEG's Ethics Guidelines.Now, let me recall another fundamental and important point about software's.Software's, and so any AI systems, only manipulates numbers (natural number in fact). Alan Turing demonstrates in 1936 that real numbers that are not calculable with a Turing Machine. There is no real numbers in a computer, only representations of them. So, data collected from human being are very "pixilated" when represented into information in a computer.Let me take a very simple example of the consequence of Turing's demonstration.It is not possible to represent an amount of money in a given currency like euro (12,55€) or dollar (12,55$) with a double precision (i.e. 64 bits) number in a computer because the arithmetic's laws (addition, multiplication, etc) will inevitably fail. Even when the data are provided as natural numbers, a simple division will render a very strong imprecision in the system (try to compute 1/3 for example ...).Another important point about numbers. All the data bases used in Deep Learning Systems are hollow, full of empty, because of the very hight dimensions of the representation space. In such a context, even the notion of average as a representation of the mathematical expectation must be questioned!As a consequence, all the questions of the AIHLEG about the notion of "data accuracy" must be formulated with the concept of uncertainty (incertitude, doubt, etc), not accuracy: any information resulting from a computation in a computer can not be accurate, except if a specific coding is used as for big numbers in astrophysical computing models or very small numbers in quantum physics models for instance.And we all know that certainty is not an operational concept to deal with responsibility: only a risk analysis holds for accountability.So, up to my humble opinion, the central question of ethics is the strength of the morphism between real world human being data and the associated information contained in databases.Imagine that a crime is proven in the digital world. What about the real world?If the strength of the morphisms is very strong, there is a great probability that the crime has been done in the real world. But if not? What if the strength of the morphisms is weak? To me, that is the fundamental question about ethics of AI systems.Another important consequence of

the representation of numbers in a computer is that it is always possible to anonymize the data: even string of characters are represented with numbers. So, the translation of strings to numbers and the reverse, numbers to strings, must be of the full and entire responsibility of the organization owning the AI system, more precisely the legal responsible of the organization. This problem is not new and AI brings no new constraints on this point.The RGPD must therefore be fully applied on AI systems as on any information system: no AI system can escape from RGPD!Finally, a point must be underlined.AIHLEG's ethics requirements can be jugged excessive by the stakeholders of AI technologies. In that case, these is a simple way to escape from these constraints: to declare that an AI system is a simple usual information system.That is why, in my mind, the term "AI system" must be defined (even if it is difficult!) in the Glossary of page iv: the qualification of AI system must be objective, not subjective to any specific or particular interests.As a minor contribution, the usual and ancient definition of an AI system can be recalled: "If a system solve a problem having no algorithmic solution, then it can be considered as an AI system". This is not a very good definition but it is sufficient for the purpose of this text.Another definition can also be used, recalling Newell's definition of "rational agent" inspired by the philosopher Daniel Dennett: "Any system acting on its environment in order to reach its goals according to a principle of rationality can be considered as an AI system". But this definition is quite more complex.A last word about what surprises me a lot.I don't understand why very important European realizations in AI are never recalled, notably the popular success of the CommonKads methodology (European project Esprit P1098 : Knowledge Based Systems Methodology Project (1985-1989)) and the extraordinary economic and technical success of the Sachem system (nowadays called "BFXpert" by Paul Wurth, Luxemburg) developed from 1990 to 1998 by Arcelor in France, the biggest AI system ever build by humans that actually equipped numerous blast furnaces over the world notably, and which has been partly founded by the UE (program ECSC-ERGONOM 6C - Sixth programme (ECSC) "Ergonomics research for the steel and coal industries", 1990-1994).Before this new century, UE founded a lot of AI research programs that are never evoked ... I don't understand why.Strange amnesia, isn't it?

Roberto V.	ZICARI	Frankfurt Big Data Lab, Goethe University Frankfurt, Germany	It would be desirable to see also a focus on the need for diversity in AI research and development teams, as this has direct impact on ethical considerations regarding AI/ML systems. If AI/ML teams are too homogeneous, the likelihood of group-think and one-dimensional perspectives rises – thereby increasing the risk of leaving the whole AI/ML project vulnerable to inherent biases and unwanted discrimination.This is something the leading AI/ML learning conferences are starting to address, and I think we can all do our part to promote the work and thinking by underrepresented groups in the research community.Examples:Women in Machine Learning (https://wimlworkshop.org/) and Black in AI (https://blackinai.github.io/) are probably good places to start.----------	Note: Feedback provided by YVONNE HOFSTETTERTwo topics are most relevant to the debate of AI in autonomous weapons, so that similar questions arise as to e.g. deployment of autonomous cars.1. Is the behavior of the autonomous system proportionate?This particularly concerns proper discrimination (combatant vs. non-combatant). Even for man, discriminating between a combatant and a civilian is difficult, because in modern conflicts combatants are no longer distinguishable from civilians. Many warring adversaries are not recognizable as fighters/soldiers because civil dressed. The decision as to which an autonomous machine's behavior is legitimate therefore needs to be made using CONTEXT. When making a similar decision, man would rely on concepts such as "good faith" or sensus communis. The philosopher Markus Gabriel speaks of the "unified impression" people have on everyday life. This is not comparable to the data salad an autonomous machine needs to fuse into a picture of a situation. Thus, with the requirement of proportionality, the question arises: Will an autonomous machine "think" like man?2. Who is accountable?In answering this question, the military is ahead of the civilian economy and industry.(1), who puts an autonomous system on the market and operates it, is accountable. If Bundeswehr orders and commissions an autonomous offensive system which was built by Airbus Defence and Space, the Bundeswehr and not the manufacturer (nor its designers, programmers, etc.), must be held liable – the manufacturer is liable (merely) according to the statutory product liability.(2), if the use of an autonomous offensive system causes damage to the civil population – such as violation-by-accident or violation-by-design –, an individual must be accountable, as a legal entity cannot be accountable (at least not according to German law).Is it then fair to blame a commander?Yes, propose some nations, but: The autonomous system must be extremely well tested. Particularly, this comprise STATISTICAL TESTS and Independent Validation and Verification IV&V. The commander must know the probability distributions when, for example, civilians may be affected in the case of deploying an autonomous system. If the commander does not know about the system's probability distribution, and intentionally or with gross negligence, uses the autonomous system anyway, and civilians become illegally affected by the system's decision making, the commander makes himself liable to prosecution (of having committed a war crime). For such a legal solution, both IHL and criminal laws would have to be amended.(3), from a legal point of view, every loss event is always a case for an insurance. In the case of whatever damage caused by (civil) autonomous systems, therefore, it will necessary to set up an insurance business, which also reflects the above mentioned thoughts.----------Note feedback provided by Steven Finlay:The question the presentation raises about regulatory frameworks is an interesting one and one should bear in mind that there is more than one perspective as to how to approach this.The EU approach (as captured by the GDPR) is very much a rights-based one. The starting point is that your data is yours and it's your right to decide how that data is	What I am missing in the draft proposal is an action plan on how to involve the key AI players.It is mandatory to involve the **key AI designers** and connect them with other relevant stakeholders.Please note that this means connecting with AI experts in USA, who are leading the field. China is also very active in this space, but I am not sure that they will be reactive to Ethics and AI...My recommendation is to contact Jeff Dean, Google Senior Fellow and SVP Google AI See this interesting video:Published on Nov 7, 2018Jeff Dean discusses the future of artificial intelligence and deep learning. This talk highlights Google research projects in healthcare, robotics, and in developing hardware to bring deep learning capability to smaller devices such as smart phones to enable solutions in remote and under-resourced locations. This talk was part of the AI in Real Life series presented by the Institute for Computational and Mathematical Engineering at Stanford University in Autumn 2018.https://www.youtube.com/watch?v=imlp8DGNkk0&index=5&list=PLn62CdVLnT-dDshwuuumF5w3rpaidb2Dm&t=0s----------I was talking with Alex Beutel (Google Brain) and this is what he pointed out as good references at Google:Google has put out some research on machine learning fairness: https://ai.google/education/responsible-ai-practices?category=fairnessA bunch of the work they are doing takes a similar approach to I think what we are suggesting in terms of trying to address concerns during model training: https://arxiv.org/abs/1707.00075 and https://arxiv.org/abs/1809.10610	Robust [and standardized?] procedures for testing and validating AIs would be a pragmatic solution, even if we don't understand fully the heuristics. Perhaps, by extensive testing with actual or synthetic data sets and extreme scenarios, an AI could be validated for its intended purpose, including likely paths of future learning?Even if we don't understand fully the heuristics, perhaps, by extensive testing with synthetics data sets and extreme scenarios, an AI could be validated for whatever purpose it is designed, including likely paths of future learning, if it is deployed in that state?In fact, I was asked a similar question when I presented the same talk at Uber in San Francisco..I thought that we do not allow kids to drive a car, they need to be at least 16 in USA, and 18 in Europe and have done a traffic school class and passed a test.Perhaps we can "certify" AIs by the number of testing with synthetics data sets and extreme scenario they went through-before allowing AIs to drive a car (similar to what happens to airplane pilots)….Somebody would need to define when good is enough. And this may be tricky…	Perhaps this useful for the discussion and the final version of the report:I recently gave a talk at UC Berkeley on the Ethical and Societal implications of Big Data and AI and what designers of intelligent systems can do to take responsibility, not only for policy makers and lawyers: http://www.odbms.org/wp-content/uploads/2018/10/Zicari.UCBerkeley.2018.pdfWe are having an interesting discussion on this topic. You can read the feedback here:http://www.odbms.org/blog/2018/10/big-data-and-ai-ethical-and-societal-implications/#comments Policy makers are actively working out legal frame for Ethical, Trustful, Transparent AI. See for example:http://www.odbms.org/blog/2018/10/on-the-future-of-ai-in-europe-interview-with-roberto-viola/ I am interested to explore how Ethics can be "embedded" into the core of the design. Not reacting to it.... Kind of "Ethics inside". We would need to talk with key AI developers to see if this is possible and meaningful and link them with policy makers and other relevant stakeholders.

				used – even if your refusal to allow use results in sub-optimal outcomes/harm. For example, by not allowing your data to be used to support medical research, others may suffer because new treatments will take longer to develop. A similar argument might be that I have a right to drive myself, even if I am less safe than an autonomous vehicle. This contrasts with the more utilitarian perspective, expressed in the quote by Steve Lohr at the start of the presentation, of thinking about data as a raw material. Data is an asset to be harvested and used. From the utilitarian perspective, one seeks to maximize the use of resources for the general good and only take specific actions to prevent mis-use; i.e. do no harm. Both a rights-based approach and a utilitarian perspective have their merits and drawbacks.The EU has gone down the rights-based approach and to date the US has been more utilitarian, but it will be interesting to see these things develop across the different regulatory regions of the world over time.----------I have found this (MIT Technology Review, Establishing an AI code of ethics will be harder than people think, October 2, 2018):“A recent study out of North Carolina State University also found that asking software engineers to read a code of ethics does nothing to change their behaviour:https://people.engr.ncsu.edu/ermurph3/papers/fse18nier.pdfPhilip Alston, an international legal scholar at NYU’s School of Law, proposes a solution to the ambiguous and unaccountable nature of ethics: reframing AI-driven consequences in terms of human rights. ”“https://www.technologyreview.com/s/612318/establishing-an-ai-code-of-ethics-will-be-harder-than-people-think/		
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Rasmus	Helt	Privat person	I agree with the "Framework for Trustworthy AI", though I would prefer in the first step of the modell at the core values to underline the need of a transparent use of data and not only a respectful one.	I agree with most points, but in "3.2 Freedom of the individual" there should also be discussed the possibility that people get the choice to say no if they don't want to be tracked by AI-systems without discrimination like for example to pay more money for insurances or other services. In "3.5 Citizens rights" there should be an "Opt in" instead of "Opt out" if data is not encrypted on highest level.	"Non Discrimination" is most important, so there should also be established a commission or another public institution like a court which checks decisions made by AI if a citizen feels discriminated.	Safety is extremily important for trust and success, so all institutions and companies should be forced to report to the public if they have noticed that data have been stolen. In many european countries like for example Germany this is not the case today and must be improved.	It is a very good idea that the European Commission starts to ask people about their opinion directly. This also strengthens democracy. Best greetings from Hamburg to Brussels.

Anonymous	Anonymous	Anonymous	We, the ACM Future of Computing Academy (FCA) also recently published a proposal (https://acm-fca.org/2018/03/29/negativeimpacts/) argueing that the computing research community needs to confront much more seriously the negative impacts of our innovations. In the words of the proposal:"The current status quo in the computing community is to frame our research by extolling its anticipated benefits to society. In other words, rose-colored glasses are the normal lenses through which we tend to view our work. This Pollyannaish perspective is present in our research papers, our applications for funding, and our industry press releases… However, one glance at the news these days reveals that focusing exclusively on the positive impacts of a new computing technology involves considering only one side of a very important story. Put simply, the negative impacts of our research are increasingly high-profile, pervasive, and damaging. Driverless vehicles and other types of automation may disrupt the careers of hundreds of millions of people [2-5]. Generated audio and video might threaten democracy [6,7]. Gig economy platforms have undermined local governments and use technology for "regulatory arbitrage" [8,9]. Crowdsourcing has been associated with (and sometimes predicated upon) sub-minimum wage pay [10,11]."To ensure that this more serious identification occurs, the FCA proposal argued for incremental changes to incentive structures in computing research, focusing on how we evaluate the quality of research reports (i.e. papers) and research proposals (e.g. grant proposals). Specifically, the proposal recommended that:"[Evaluators as well as companies of research should] require that papers and proposals rigorously consider all reasonable broader impacts, both positive and negative."The proposal contended that this is not only a matter of social responsibility, but also one of intellectual rigor. A proposal that motivates its research with its positive implications but does not discuss any of its negative implications is providing a picture of the corresponding research that is incomplete in important ways. Given that it is the job of research evaluators to ensure that papers provide a complete picture of the described research, the FCA proposal argued that adopting its recommendation can be understood as already existing within the evaluator mandate.In addition providing a transparent description of potential negative impacts, the proposal also argued that authors should be encouraged to discuss means by which any potential negative impacts might be mitigated. This discussion might include a description of further research, new regulations or other approaches. This would then make it much easier for those who seek to execute that important follow-on research, or develop that policy, to motivate their efforts.We believe that this perspective should be included in the guidelines as well.			I would like to express my feedback on the Draft AI Ethics Guidelines for Trustworthy AI, prepared by the High-Level Expert Group on Artificial Intelligence. I am a Lichtenberg Professor and Professor of Human-Computer Interaction (HCI) at the University of Bremen in Germany. In addition I am the co-director of the Bremen Spatial Cognition Center (BSCC) and member of the TZI (Technologie-Zentrum Informatik und Informationstechnik) and Minds, Media, Machines (MMM). MMM is an interdisciplinary network of researchers at University Bremen, Germany.My research interests lie at the intersection between (HCI), geographic information science and ubiquitous interface technologies. In our HCI lab we investigate how people interact with digital spatial information and create new methods and novel interfaces to help people interact with spatial information. This includes the development and evaluation of wearable technologies, mobile augmented reality and virtual reality applications, interactive surfaces and tabletops, and other "post desktop" interfaces. www.johannesschoening.deMy research and work has received several awards, such as the ACM Eugene Lawler Award, a Vodafone Research Award, the lasting impact award at MobileHCI and two Google Research Awards. I regularly talk about the impact novel technologies and consult with companies and thinktanks. In addition, I serve as a junior fellow of "Gesellschaft für Informatik" and I am a member of the ACM Future of Computing Academy.* TimingIt is not optimal to have such an important consolation within a month. I would strongly encourage to have multiple iterations and all all stakeholders to have more time to engage with those important guidelines. * RegulationIn addition those guidelines are not enough we need good legislation and regulations to make sure technology is developed for the good of all and not for a small minority. Guidelines are a start, but not enough. *Being Flexible The guidelines as well as the related legislation should be flexible enough to cope with the rapid changes of research and technologies in the field of AI. What is the process, that the guidelines are updated? How to react to rapid changes? *Industry InvolvementIt is unclear how industry helped to shape this first draft on the AI Ethics. The process itself needs to be transparent and not biased towards certain stakeholders.
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential

Gabriele	Sprengseis	Österreichischer Behindertenrat			Page 16, Non-Discrimination: In the case of the types of discrimination, always also call disability, not just ethnicity, gender, sexual orientation and age. Disability should be enumerated so that it is visible. Page 20: Testing & Validating And last but not least, everything should be tested for accessibility. Page 22: "... that Teams are diverse in terms of gender, culture, age and Disability, but also in terms ...		
Piotr	Jędrzejowicz	Gdynia Maritime University			Trustworthiness is a fuzzy concept. While, in general, we try to develop trustworthy systems, what is trustworthy in one area of applications might not be trustworthy in another. One of the main factors of trustworthiness is reliability. One has to accept that it is not possible to construct AI systems that are fully immune to failures.	A lot of research is needed before we will be able to assess the trustworthiness of the AI systems in a way that is measurable and easy to understand by persons outside the field.	In my view, the Guidelines, in its present form, are too general and too idealistic. Besides, they can be reasonably applied to all kinds of technical systems.
Anonymous	Anonymous	Anonymous	First, I would like to commend the AI HLEG's work on these guidelines and the introduction in particular. Listing all the positive points in this document would by far exceed the scope of this feedback form. As such, I will limit myself to the few points I would recommend to change.* In the glossary preceding the document, the definition of 'ethical purpose' remains vague and thus may be insufficient to give the reader an understanding of the kinds of core principles and values the document is referring to. In order to keep the glossary entry short it may be helpful to point to later pages in the document where these principles and values are discussed in more detail.* In 'The Role of AI Ethics', the document provides a definition of AI ethics, but, in my opinion, does not clearly connect this definition to the guidelines as such. This point could be addressed by rephrasing the second paragraph and connecting it better to the first paragraph (e.g. by making clearer that the 'ethical reasoning' mentioned refers to the first paragraph)* The sentence 'We therefore assert that our European AI Ethics Guidelines should be read as a starting point for the debate on Trustworthy AI' appears redundant, given that the previous and following sentences seem to have the same meaning.	Paragraph 3.4 states that 'Equality means equal treatment of all human beings, regardless of whether they are in a similar situation.' I support this definition and highly appreciate that the AI HLEG has decided to promote a more demanding understand of equality than mere non-discrimination. However, I am not yet certain that the remainder of the document mirrors this understanding of equality sufficiently. Indeed, the term 'equality' is not used at any later point in the document and is instead replaced by alternative notions like 'fairness', 'equal opportunity', or 'equal treatment'. If this is the understanding of equality in the AI HLEG's terms, then this is absolutely fine. However, I would recommend to make this understanding of equality transparent such that the lack of the term in the rest of the document is more understandable.In the list of critical concerns, the document does not list personalization, which may also be critical, especially if personalization entails differences in opportunity. An example is the long-standing policy of special needs schools in Germany for people with disabilities, which in many cases provided an education of poorer quality compared to other school forms. This fact has been (rightfully) criticized in light of the UN inclusion requirements. A more benign (yet more widespread) example is personalized pricing. Personalization also may be critical from a democracy and citizenship perspective in light of filter bubble concerns.The section on Lethal Autonomous Weapon Systems does not include the concern that such systems potentially lower the entry into armed combat. To date, the risk of having soldiers injured or killed in combat poses a deterrent for engaging in armed conflict. If this deterrent is removed, governments (or private parties, for that matter) may be more inclined to use lethal force instead of diplomatic means. As such, LAWS may, paradoxically, increase the death toll of armed conflicts instead of lowering it.The paragraph on longer-term concerns, as it is now, limits itself to long-term concerns that are highly unlikely or theoretical. In doing so, the paragraph fails to mention a multitude of long-term concerns that seem, at least to me, both more likely and more dangerous. Two particular kinds of long-term	It remains unclear whether the list of requirements specified at the beginning of the chapter constitutes a sufficient set of requirements to comply with all the principles and values specified in chapter one. Maybe the AI HLEG could make this clearer.Further, while I understand the reasoning behind an alphabetic ordering of the requirement list, I would recommend an ordering that facilitates understandability a bit more. In particular, I would recommend to re-structure the requirements list such that requirements which have been derived from the same ethical principle/value are grouped together.If possible, it may also be helpful to use a bullet point list instead of a numbered list to visually express that all points are equally important.Section 2 on data governments suggests to 'prune biases' from a dataset. While I appreciate that this is promoted, it may not always be possible to prune bias, and instead one must apply different means to train a system that is unbiased from the biased data. Of course, if the AI HLEG finds that such alternative methods can, in principle, not succeed, this is a valid point, but should maybe made more explicit.Further, I would suggest to add a sentence regarding the updating of data sets. In many cases, biases may occur because the world is not stationary and thus training data recorded some years ago may not apply to the current situation. In such cases, data sets have to be augmented or replaced with current data to ensure that a system is still applicable.Section 3 on 'Design for all' lists 'age, disability status or social status' as dimensions of inclusion. I wonder whether the limitation to these three dimensions is intentional, given that the EU charter of fundamental rights lists many more dimensions of potential discrimination. Otherwise, it may be viable to either extend the list or use an umbrella term. Finally, I note that equality/fairness/justice concerns in the requirements list seem to be limited to design for all and non-discrimination, which leaves other dimensions of equality to be desired, in particular equality in terms of harms and benefits.In general, the list seems to be focused mostly on doing no harm instead of actively doing good.The section on 'Architectures for Trustworthy AI' demands that 'ethical goals and	Point 2 on data governance lists 'Is an oversight mechanism put in place? Who is ultimately responsible?' These questions seem to be far more general than just data governance. If these questions are meant to be specific to data governance, they maybe should be rephrased. Further, I would recommend to include a question similar to 'Does the data include sufficient variability to represent the full space of situations in which the system should operate?'In line with my comments on the previous chapter, the assessment list at present does not seem to include points regarding equality issues beyond non-discrimination or design for all. Including such issues would be appreciated.Regarding the use cases, I have the following comments. Note that these do not exhaustively cover the use cases but are, rather, constrained to my areas of expertise.Healthcare Diagnose and Treatment:* Regarding data governance, medical data is particularly sensitive and should be stored encrypted and safely. If data is released for scientific purposes, it should be ensured that anonymization is ensured and de-anonymization is impossible.* Regarding data governance and robustness, training, testing, and validation must be performed on data that is sufficiently variable to include a wide range of people, especially including people with side conditions that may complicate diagnosis* Regarding both design for all and non-discrimination, the system must ensure that diagnostic accuracy is equal across all demographics which may be subject to the condition in question (e.g. the diagnostic should not be more accurate for men compared to women)* If there do exist specific cases where the system is not applicable, e.g. due to the presence of conditions that overshadow any symptoms of the condition in question, this must be made transparent. Further, a fall-back plan must exist which permits to properly diagnose and treat these cases independent of the system* Regarding transparency, any lack of accuracy, especially in terms of false positives, must be made transparent; it must also be considered that the ability to give informed consent is complicated by the fact that people may not be aware how they react to a falsely positive diagnosis with a	Again, I would like to extend my compliments to the AI HLEG to having drafted a very helpful document that builds well upon existing research and has accumulated a dense and well-structured set of guidelines.To make the structure even more clear, I would recommend to provide another diagram which visually connects the five ethical principles from chapter I with the ten requirements in chapter II and the technical and non-technical methods in chapter II, i.e. a variant of Figure 1, but with the single terms filled in.Further, as a very minor point, the figures throughout the document are pixel graphics but could be replaced by vector graphic versions which would facilitate printing in high resolution. Ideally, the AI HLEG could also release these figures (and the entire document) under a license that facilitates re-use such that it can easily be used for teaching purposes (such as Creative Commons).Finally, in addition to the glossary in the beginning of the document, it may be worthwhile to have a list of terms at the end of the document, accompanied by pointers to the page numbers where these terms are explained. This would be particularly helpful as the document introduces quite a lot of terms (e.g. auditability or explicability), not all of which are intuitively clear.

outcomes come to mind:First, negative effects due to feedback loops in the application of AI decision making systems. Such feedback loops may drive speculative bubbles in algorithmic trading, overpolicing of minorities due to predictive policing, gender stereotyping due to personalization, and so on. In general, undesirable feedback loops may occur whenever the predictions of an AI system influence human decisions such that the prediction becomes more likely (self-fulfilling prophecy effect; also refer to O'Neils 'Weapons of Math Destruction'). This effect exists for human decisions as well, of course, but AI systems can provide these systems on much larger scale, at much higher speed, and under the guise of objectivity.Second, negative effects due to conditions which lie outside the training data set for the system. In most cases, AI systems which have been trained on a sufficiently large training data set operate predictably. However, there exist circumstances which lie far enough outside the set of data on which the system has been trained, tested, and validated, that the system will behave in a highly unpredictable and potentially dangerous manner. This is of particular concern in systems that could take actions which endanger human lives (e.g. infrastructure, health care systems, autonomous vehicles). Note that this is similar to the 'black swan' concept but may connect better to human experience as we have witnessed many instances in which AI systems made wrong decisions under such circumstances (as an example, one may consider the infamous 1983 Soviet nuclear false alarm incident in which an early AI nuclear early-warning system may have lead to a nuclear war if not for human intervention).I believe that these two long-term concerns may make a stronger argument compared to concerns regarding artificial consciousness.

requirements should be integrated at "sense"-level' of an agent. This phrasing seems to imply that, if ethical goals and requirements are implemented at this stage, all subsequent reasoning and acting is automatically ensured to be ethically fine. If the AI HLEG does not which to make this implication, a slight rephrasing may be in order.The list of non-technical measures does not include risk analysis. This may be included in the stakeholder and social dialogue. If so, maybe it could be mentioned there.

serious condition. This has been discussed e.g. with respective to population screenings for rare conditions where false positive rates are high* In addition to the requirement list, healthcare diagnostics and treatment should aim at promoting human health and well-being beyond what could be achieved before; this is a rather obvious point but should be emphasized in light of ethical purposeAutonomous Driving/Moving:* Regarding data governance and robustness, the data must include a very wide range of driving conditions, including weather variability, lighting variability, highway versus country versus city traffic, etc. Further, these conditions should be made transparent such that audits can take place and, in case of faulty behavior, the problem can be traced back to limitations of the training data* Regarding non-discrimination, sensing and acting should behave equitably across demographics, i.e. a car should be able to recognize any other traffic participant independent of their apparent features and should act to promote their safety.* Beyond the current requirement list, the autonomous vehicle should aim at doing good by reducing the need for fuel and other resources whenever possibleInsurance premiums:Insurance premiums pose a particular challenge because issues of AI are interlinked with broader societal issues. I will try to limit my comments to the application of AI systems in particular; yet, connections to other issues are not entirely avoidable.* Regarding non-discriminations, risk variables are very likely to correlate with dimensions of discrimination, e.g. gender, age, etc. While these correlations do exist, system designers should ensure that decision variability is due to the root causes for risk, not due to group variables (gender, age, etc.) that are merely correlated, but not causally linked* Further, systems must be designed to not induce self-fulfilling prophecies. For example, if an insurance premium is high, fewer resource are available to prevent a risk, which poses the hazard of ever-rising premiums interlinked with ever-increasing risk* Regarding design for all, high insurance premiums necessarily exclude poor people; depending on the kind of insurance, it may be necessary to implement methods to still grant poor people these kinds of insurance (fall-back option) or to prevent systems from charging high premiums to people who can not afford them* Regarding accountability, autonomy, and robustness: The insurance company must provide ways to appeal decisions regarding high premiums or increased premiums and respond to an appeal timely (or reduce the premium until the appeal is addressed)* Regarding transparency, the system should offer explanations, at least a list of features and their influence, that lead to a high premium decision; it would also be possible to provide counterfactuals, e.g. examples how the person could be different to achieve a lower premium; these counterfactuals should be actionable for the user* Regarding autonomy, insurance premiums should not punish risky behavior unduly, and ideally should promote behavioral diversity, i.e. it should not limit low premiums to a set 'ideal' lifestyle as implicitly or explicitly defined by the systemProfiling and law enforcement:* In profiling and law enforcement, particular care must be taken to not cause feedback

						loops with undesirable outcomes; for example, if the system suggests higher police attention to a certain group/region that is deemed higher-risk, this increases the chance of noticing crime in that group/region, thus further increasing the risk score, and so on; note that these effects exist with or without AI systems, but that AI systems can potentially speed up such cycles and worsen their impact* Regarding data governance and robustness, exceeding care must be taken to update data and ensure that enough exploration takes place to spot data that is inconsistent with previous assumptions; crime is highly dynamic and tends to adapt to law enforcement strategy, such that any system that is overly rigid will fail to address it accurately* Regarding accountability, autonomy, and transparency, citizens must have the option to question the reasons why they were identified as a risk by the system and have the right to appeal that identification, even if the decision turns out to be correct; in other words, a risk scoring system may have identified risk correctly, but for the wrong reasons, in which case an appeal should still be possible* Regarding non-discrimination, care should be taken to apply such system across the board and not exempt certain groups from automatized treatment, e.g. white collar crime* Regarding data governance, a risk assessment by an automatized system should not be taken as evidence that a person is guilty as such; instead, a legal process should inspect the reasoning behind the assessment and check its soundness before rendering it viable for any further legal use	
Jonathan	Sinclair	Celgene	On page ii, open reference is made to the fact that: "These Guidelines are not meant to stifle AI innovation in Europe, but instead aim to use ethics as inspiration to develop a unique brand of AI, one that aims at protecting and benefiting both individuals and the common good.". This will lead to the generation of adversarial AI agents, opposed to what Europe considers the 'common good'. Leading to a concern that despite the EU's lauded position to human rights and its success at exporting these ideals globally, countries will exist that will exploit the EU concept of a “common good” to gain an advantage. This may lead to an innovative market for EU branded AI that is globally desired, however experience with globalization may tell us that this will restrict innovation. We've seen this with the US vs. EU engagement in technology, cloud adoption, etc. and US vs. China in production. There is valid reason for concern that the establishment of EU AI, that is culturally rooted in the ideas of the EU, will only be fit for the EU. This may not be negative in totality, but it is perhaps naive to say that it won't stifle innovation. The advancement of self-driving cars in the US (without ethical consideration) could be presented as an example.	Anchoring AI to fundamental rights is an interesting idea, although when one tracks the successes of enforcing these principles one is left wanting, the Iraq and Ukraine incidents are two such cases where, in general the world lauds and subscribes to these notions, however in reality the effectiveness of enforcement is largely ineffective. This being the case, it's hard to see how aligning to these principles will help drive a strategic ethical direction of AI. There also appears to be a clear contradiction in "The Principle of Beneficence: “Do Good”" and "The Principle of Non maleficence: “Do no Harm”". The first focuses on: "improve individual and collective wellbeing" going as far as to say that 'do good' is equal to: "generating prosperity, value creation and wealth maximization and sustainability". If an explicit subscription to capitalist ideals makes up the principle of beneficence there is a fear that minorities could be side-lined. This is a massive issue as most minorities and human rights principles are not aimed at the masses, but at repressed sub-cultures. An AI may freely determine that it is doing good by maximizing general prosperity, but maximization can't occur universally, generating a conflict with the principle of non-maleficence: "Harms can be physical,	This section seems confused with topic duplication between items such as requirements 3 & 5, 7. Item 1 seems too weak. It should go further referring to an EU review board and revision of local laws. There is a question whether item 2 could ever be technically implemented. The inclusion of the following: "Diversity and inclusive design teams" under the 'Non-Technical Methods' section is interesting. The document is focusing on an EU code of ethics, based on EU standards and yet there is a mention of ‘diversity and inclusion’, when in fact, what is wanted here is not really diversity, but alignment to EU ideals. Section 5 seems very valuable but not developed enough particularly 5.4. I'm inclined to disagree with points advocated by the likes of Prof. Sharkey and others, however an alignment needs to be found between LAWS and EU's article 3 of TEU. This part seems to have been added, for the sake of inclusion, rather than providing anything meaningful.	This section is simply a duplication of the previous with more context added to help understand the principle requirements. There are few new points added here and the addition of questions removes any ability to provide guidance except in a few places. The rationale for this section should have posed the questions and provided a response according to the EU position e.g. "Has an Ethical AI review board been established?". If yes, where can the ethical review be located, if no. The AI will not be made available to the community and could be shutdown on based on infractions to the EU directive. (As an example). Alternatively more concrete statements should have been issued e.g. in Part 4: Governing AI autonomy: the following "Is a "stop button" foreseen in case of self-learning AI approaches?" should be replaced with "A stop mechanism must be developed to ensure that should an self-learning system start deviating from its ethical mandate the system can be safely disabled in accordance with EU Ethical AI directives"	The document is an interesting idea and it's good that this is being discussed. Establishing the ethics of AI based on existing EU principles of human rights protection makes a lot of sense, however the document loses its way in many places. Consistency isn't followed, and this is perhaps reflected with the diverse number of interested parties that went into its initial creation. I would have preferred to have seen a clearer focus of EU driven direction and concrete positions. It seems that private interest groups and academics have pushed their respective agendas which has resulted in a dilution of the documents overall value.

				psychological, financial or social". Arguments of this type have been worked and re-worked, largely stemming from Asimov's laws. There is a feeling that this section has avoided the current state-of-art in these discussions. Fairness and goodness should be dropped with an increased focus on the sections of Autonomy, Excitability and Accountability			
Marius	Schulz	private		I want to refer to an article showing a distopian scenario, but arguing on ethic priciples of AI somehow and especially in possible connection to nano robots may developed: https://marius-a-schulz.de/2018/12/04/spieltheorie-nano-roboter-gehirnbeeintraechtigung/ .There is a translation widget in the menu.Regards,M. Schulz.			
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Alan	BUNDY	School of informatics, University of Edinburgh					The main omission in these guidelines is any discussion of the need for better public understanding of AI. Recent successful AI applications have provided World-class performance in a very narrow area. Such 'intelligence' is outside the experience of most people. There is a danger that they will overestimate the capabilities of such systems by trusting them beyond their capacity. I've previously written about this danger in Smart Machines are Not a Threat to Humanity, Bundy, A. Feb 2017 In Communications of the ACM. 60, 2, p. 40-42. Additionally, unrealistic expectations about the capabilities of AI systems can arise if you don't understand how they work. These dangers touch on several of the issues raised by AI HLEG. 1. Transparency: The 10th requirement for trustworthiness states "Explainability – as a form of transparency – entails the capability to describe, inspect and reproduce the mechanisms through which AI systems make decisions …". But in most statistical machine learning systems, such explainability is not obtainable: their mechanisms are inherently opaque. This is briefly alluded to on p20 "A known issue with learning systems based on neural nets is the difficulty to provide clear reasons for the interpretations and decisions of the system". However, elsewhere the guidelines refer to "the causality of the algorithmic decision-making process", "AI systems should document both the decisions they make and the whole process that yielded the decisions, to make decisions traceable", etc, even though such traces are not always obtainable. One wonders whether the lawyers who drew up the GDPR understood that they were asking for the unobtainable. 2. Judgement: Section 5.4 on LAWS says, "in an armed conflict LAWS can reduce collateral damage". This makes unrealistic assumptions about the capabilities of current autonomous systems to distinguish, for instance, combatants from non-combatants, especially when informal forces are involved.

						Failure to make such distinctions breaches the Geneva Convention. 3. Verification: The 9th requirement for trustworthiness states "Moreover, formal mechanisms are needed to measure and guide the adaptability of AI systems". I'm an advocate of formal methods, but they are not a panacea. Take the case of neural nets. Verifying that a neural net tool satisfies a specification is feasible, but that does not apply to the classification application that is obtained by using the tool to train a network on examples and non-examples. The applications correctness depends, for instance, on the representativeness and size of the training set. Biased training sets will lead to incorrect behaviour even using a verified neural net tool. Overfitting will occur if there too many features are used given the size of the training set.
Lars	Jaeger	Independent Writer and Author (www.larsjaeger.ch)	The document starts on some bad footing, as it enthusiastically describes the great potential benefits of AI (which undoubtedly exist), while at the same time uses some very nebulous and defensive language when it comes to its risks. The sentence in the Executive Summary "AI also gives rise to certain risks that should be properly managed." is way too vague and seems to imply that the risks are always and necessarily controllable "if just properly managed". We do not know if that is true. We can easily imagine developments in which entirely new structural complexity in AI systems appears very suddenly and unexpectedly and leave us without the ability to "properly manage". I suggest the sentence above to be modified to "AI also gives rise to certain risks, some of them of existential nature that should be properly managed. The expert group is aware, however, that there exist "unknown unknowns" in the possible features of a future AI technology the possibilities of which we should point particular attention to." In general, low likelihood developments with devastating consequences (possibly related to Unsupervised Recursively Self-Improving Artificial General Intelligence, AGI) always have to remain a strong focus of society when it comes to AI.	In the context of AI risks, Section I.5.5 eludes to some of the longer term risks of AI but remains especially vague. It appears that this was the section that least agreement was found on within the expert group. However, this does not mean it is not important and should not be broadened (in fact it is rather a reason to do so). At least any reference to "a very distant future" in section I.5.5 should be completely avoided, as we all know that the distance future might not be that distant when it comes to technological developments. Another important point of criticism deals with the principle of "explicability" of effects supplied by AI, as it is being outlined in Section I.4, last bullet point. The hiding behind this demand for explicability, which is to be achieved by users giving an "informed consent", might be interpreted as a desired backdoor for AI developers to undermine fundamental principles of human dignity, for example, by pretending explicability to the user in order to get his consent. If the user then presses the "Okay" button, the AI developer would not only be legally but also ethically off the hook, as, after all, the user has agreed to the explanations. If one considers the implementation of the General Data Protection Regulation (GDPR) by Google, Facebook and Co., such fears can hardly be dismissed. Why is the criterion of explicability necessarily and solely informed consent by the user? The next sentence "Explicability also requires accountability measures be put in place" remains very nebulous and does not remedy the issue. We should be a little more demanding on AI developers and AI product providers. The outline in the first paragraph of that bullet point "Technological transparency implies that AI systems be auditable, comprehensible and intelligible by human beings at varying levels of comprehension and expertise. Business model transparency means that human beings are knowingly informed of the intention of developers and technology implementers of AI systems. equally remains too vague that purpose. It is easy to imagine how AI companies use such description to comfort users rather than give them enough information about the implications their AI algorithms have on them and their lives.		The discussion about the design of AI technology is surely too important to be solely guided by the capitalist logic of exploitation and led only with regards to the return prospects of tech investors. It must be conducted on the basis of broad democratic processes in which a broad spectrum of interests and opinions are involved. Therefore, the AI HLEG's paper on Ethics Guidelines for A Trustworthy AI is to be welcomed in principle, as it provides a timely and well-structured guideline of how to develop a technology to our benefits rather than harm, which as the document states correctly "is one of the most transformative forces of our time, and is bound to alter the fabric of society". Giving my general endorsement of the document, I believe it nevertheless gives rise to some point of justified criticism as well as some ambiguities in wording. Also, on a general note: The time window for discussion is quite tight, plus falls right into the holiday season. This might lead to the suspicion that a broader discussion may be not quite as much wanted as expressed upon release.

				Under “Ethical Principles” in Section I.4., third paragraph, the document refers to the EU Charter, in which fundamental human rights are explicitly set out. The context of that reference is “In particular situations, tensions may arise between the principles when considered from the point of view of an individual compared with the point of view of society, and vice versa.” In such contexts of ambiguity, the document suggest, “it may however help to return to the principles and overarching values and rights protected by the EU Treaties and Charter ". One would wish for this to be expressed a little more clearly. First, it may not only help but be deemed vital and necessary to adhere to the principles of the EU treaties. Secondly, such adherence is unconditional to any situation. There simply should not be any doubt or ambiguity over adherence to the principles of human dignity and other fundamental values laid down in the EU Treaties.		
Patrik	Floréen	University of Helsinki	The basis for the document is sound.	Concerning harm and bias, it is not always clear what is bias and what is not. For instance, some handicap should in some cases be a decisive factor, e.g., when deciding about social welfare intended for the handicapped. So one cannot say that a certain trait should never be decisive, but this is still an easy case. What about statistical differences, can they be used to what extent? For instance, should insurance premiums be allowed to be higher for persons with many diseases? Or for people in certain areas of the country, where it is well-known that people on average are less healthy? We can refer to the laws about non-discrimination, but they are not necessary the same in all EU countries or interpreted the same way (example: same-sex marriage). Concerning the right to opt out from interaction with AI systems or the right not to be subject to decisions by AI, I think this is not necessary. There are many routine decisions that are fully rule-based and there is no reason to limit the use of software robots to do them. For instance, the amount of tax to pay has no subjectivity. Entering the numbers, you get the tax. The phrase on page 10 “positives and negatives resulting from AI should be evenly distributed” means what? As it is written, it could mean that half of loan applicants should be granted the loan, half not. This apparently is not what is intended, so this would be good to rephrase. In Section 5.1. on page 11, there is a passage about “identification using biometric data” with explanation in parenthesis “personality assessment through micro expressions”. But personality assessment is not identification. In Section 5.2. on page 11 “robots that are built to be as human-like as possible”: they are not built so, cf. uncanny valley. In Section 5.3. on page 12 it is said that users should have the possibility to opt-out from a scoring mechanism. So if there is some automatized system of points from traffic offences, resulting in a fine or other penalties, you say the citizen should have	Some points are conceptually very close to each other and it could be a good idea to merge them. Design for all and Non-discrimination are very close and in the following part, assessment list questions about fairness are included for Design for all, while fairness is here mentioned in the context of Non-discrimination. Also the points on Respect for Human Autonomy and Respect for Privacy could be joined. A shorter list is better. Mid-page 15, what means the concept “horizontal category of society”? It is very good that it is mentioned that AI can be used to identify inherent bias in existing human decision making (page 16). On page 20 there is a section on Traceability & Auditability. Traceability is a problem in deep learning. I hope you do not aim at forbidding use of deep learning. The next section about XAI seems to be about essentially the same thing, so that is confusing. In the previous part, the word explicability was used, but not here or in the following part of the document. The last sentence on page 20 was not comprehensible (“undergo a digital transformation”?).	Question set 1: Why would it be important to have staff of different background working on AI? So you want to have in each AI developer team some children, some handicapped etc.? It is, however, important that different minority groups are taken into account in the design. On page 8 it says that presence of an ethical expert is advised to accompany the design, development and deployment of AI. This seems to be missing from the list. Question set 3/5: It is important to simulate persons from different groups: how would this system be used by elderly, handicapped, etc. Explicability could be mentioned explicitly. Now there is a mention in passing about users “having the facility to interrogate algorithmic decisions in order to fully understand their …” I disagree about the “fully understand”. People do not ask how google work when they google, and they need not know the technical details about what is happening. The same for their car navigation system. Do they need to know how GPS works and how the AI calculates the fastest route? They don’t. They only need a vague understanding of what is going on.	Good report, more concise is better.

				the power to say that they do not want to be part of such a system? I think they should not have the option to opt out. Please do not go too much into science fiction (AGI). Good that the section was short.	
Anonymous	Anonymous	Anonymous		5. Critical concerns raised by AI5.5 Potential longer-term concerns I support efforts for the reduction of uncertainty on this matter. Concretely, it would seem valuable to support & bring together existing efforts to monitor progress in AI. This inter alia includes the Electronic Frontier Foundation (https://www.eff.org/ai/metrics), AI Impacts (https://aiimpacts.org/), AI Index (http://cdn.aiindex.org/2017-report.pdf) & Deutsches Observatorium für Künstliche Intelligenz (https://www.bmbf.de/files/Nationale_KI-Strategie.pdf#page=26). Subsequently, the HLGAI should strongly consider recommending that the EU joins the initiative by France and Canada to build up something like an "IPCC for AI".(https://pm.gc.ca/eng/news/2018/12/06/mandate-international-panel-artificial-intelligence) that can help to monitor and deal with social consequences of AI as well as scan the horizon and give early warnings if we get closer to the types of AI mentioned in this section.	2. Technical and Non-Technical Methods to achieve Trustworthy AI 2. Non-Technical MethodsControl Access to Certain Types of Technology / ResearchThis whole document seems to be based on the assumption that only responsible agents and organizations will develop and use AI systems, however, that's just not realistic. The human population contains about 4% sociopaths and 1% psychopaths. Furthermore, criminal gangs have historically often been early adopters of new technologies. As pointed out inter alia by a group of researchers in the report "The Malicious Use of AI" there are plenty of ways in which actors with malicious intent could leverage AI tools.https://arxiv.org/pdf/1802.07228.pdfA current real-world example of malicious use of AI is „deepfake porn", where the face of people is matched onto pornographic movies nonconsensually. Most of it has been created using Google's openly available TensorFlow. See: https://qz.com/1199850/google-gave-the-world-powerful-open-source-ai-tools-and-the-world-made-porn-with-them/So, Google can have the greatest ethical standards ever, but if their tools are open for anyone without appropriate measures of control, they will be used for everything. Therefore, part of the ethical guidelines should be that ethical actors have to make efforts to ensure that their technology and research does not fall into the wrong hands. This can start with very basic things such as people having to register with official documents before getting access to certain tools.We have developed systems of controlling and restricting access to research, materials etc. for any powerful technology that can be misused, including for example nuclear research, conventional explosives, firearms, research on biological pathogens or conventional computer malware. There is simply no way around this for AI tools as they get more and more powerful. Patents, Algorithms and Data TrustEthical principles are nice, but unfortunately they often have a limited impact. Many Tech CEOs will happily endorse some high-level abstract aspirational principles and not change their behavior in any meaningful way. At best companies will just create some jobs for a toothless ethics committee. That may be a bit overly cynical, but it highlights the game theoretical problem of a non-regulatory approach.I'm convinced that bottom-up ethics by a coalition of the willing can be effective. However, only if that coalition uses the right tools to make cooperation stable and leverage their power. Specifically, I have the innovative use of the legal form of trusts in	

					mind that can create something in between the current open-source vs. privatized paradigm. Donations to a trust are irreversible and a trust is legally required to only use donated assets in accordance with its stated mission. If the EU creates a trust that aims to use AI for the purposes stated in these guidelines and that has trustworthy governance, researchers, firms or individuals can donate things such as patents or personal data to it. The trust can then ensure that only firms / individuals etc. that comply with a certain set of ethics get access to this / profit from it. Needless to say that it will be harder to win over tech CEOs for this than just for signing up to abstract principles, but that's the difference between hot air and a trustworthy commitment to ethical principles.		
Stefan	Bergheim	Center for Societal Progress / Zentrum für gesellschaftlichen Fortschritt	As an organization focusing on quality of life and wellbeing, we highly appreciate that the guidelines use those concepts as starting points. However, the guidelines unfortunately do not use the insights from these research areas. There is a huge literature on human needs (e.g. relatedness, recognition, orientation etc.) that the guidelines do not refer to. There have been major efforts around the global to operationalize wellbeing, such as the OECD's Better Life Index. The guidelines mention "grand challenges" but fail to go step by step through at least one of these challenges to show how exactly AI contributes to wellbeing. We are working on a dialogue process on quality of life in the digital era that links the two debates. www.gutlebendigital.de/topics?lang=en In general, too high a hope is based on AI in the document (page i: "is key for"). Other human technologies such as the will to reduced CO2 with a well-working carbon tax might be much more important.	The rights' based approach seems difficult to operationalize. As mentioned, we would suggest a needs' based approach. The five principles do not appear to be fully intuitive and might benefit from more background than the medical context (is that applicable here?).		It seems that a lot of crucial ideas from the earlier chapters are lost on the way. Where did the "do good" and the "do no harm" principle go? Maybe we missed the links into chapter III.	There appear to be several parallels to the situation in the financial industry some 15-20 years ago. "Trust" was also a key topic there, but that did not help in the end. The breakdown came from unexpected angles. What we would suggest in the case of AI is an open scenario process in the spirit of Peter Schwartz, that would focus on key uncertainties and develop four highly diverse scenarios. This could make visible some more of the key issues to be dealt with. Two root problems in the financial industry were the huge political will in the USA to have housing ownership for as many people as possible and the largely self-regulating stance of the industry. There are parallels to the pro-AI political will now and to the codes of conduct of corporates. We again have good-sounding codes, but the incentive structures within large organizations suggest other actions - as in finance. Also, we need truly independent watchdogs (re Finance Watch) and a big-picture view of systemic risks rather than getting lost in debates on details.
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Anonymous	Anonymous	Anonymous					It has to be able to feel the sense of wonder (which presumably comes from favorable conditions) to avoid trying to destroy other life. It has to be aware of the deterministic nature of reality in order to understand that no one is fundamentally responsible for their actions and therefore should not be punished. It has to be aware of the trajectory of intellectual progress, so that it doesn't think that a solution to kill everyone is ok because that seems like the only option NOW. More intellectual development should prioritized.

Anonymous	Anonymous	Anonymous	A proper working definition of AI (re. page IV, def. of AI): Part of the reason that AI is hard to define, is that it does not (in a sense) differ from any other automation-based innovation, and fundamental ideas underpinning AI are around since (at least) the 1930s. The new context of available storage and computation however does affect this. So, a reasonable working definition (for the sake of this document) could be 'An AI driven technology has a component of data-intensity, computation-intensity, adaptivity (personalisation) and/or technical autonomy (or any mixture of those)'. Note that these concepts are 'relative' at best (that is, what was computation-intensive in the 1970s is not necessarily so in the 2010s).	A poignant starting point to exam any solution would be to ask the question: 'Should it be allowed to have a superior AI-empowered pacemaker to enter the open market?'	Robustness and reliability (p 17, chapter II): 'Trustworthy' is translated into 'robustness' and 'reliability'. One can argue that this is only partially covering the intended idea: 'robustness' needs one to specify 'x' in 'robust against x' (robust against outliers, robust against malicious errors in the training data, robust against misspecification, etc.), and 'reliability' concerns (mechanical) failure. However a third leg is phrased too implicitly in the current document: namely the concern that legitimate use of a certain technology should also work well in (future, unforeseen) circumstances, or that governing instances should have the ability to de-activate certification in case of future unintended consequences. In science, theoretical results are always conditioned on their stated assumptions (limitations). Absolute statements (absolute certifications) are impossible, and the best one can do is to make the assumptions sufficiently broad to cover intended use. Academical proof is then (essentially) stringing assumptions together into a new result. The government-based practice of certification should reflect this: we suggest to introduce the idea of 'smart certificates' were activation is a continued effort (re. Introduction to Section II.2, p. 18). Legitimate use of an AI solution would then be based on the activation of the 'smart certificates' covering all its ingredients. The task of governing institutions is to maintain unbiased evidence for the 'smart certificates' which the developer produces.	Digital twin: A constructive solution for governing AI tools (re. Section 2) is to introduce the notion of a 'digital twin'. This concept is finding traction in the IT-security arena. A digital twin collects all (digital) information of its respective human user, and serves questions from the external IT sphere. All interaction of the human user with an AI solution should go through her/his digital twin, hence serving as his/her gateway (or firewall). The use of such scheme (1) shields human users from the technical complexities, (2) provides a clear pointer to where she/he can overview all her/his digital traces, and (3) engages her/him to take personal responsibility where needed. This gateway is de-facto implemented nowadays by the system of 'corporate' smart phones, feeding corporate agendas etc. A proper implementation of the idea of the digital twin will lead to a clear separation of responsibilities.

JORGE PEGADOLIZ EESC

PRINCIPAUX DOMAINES ÉTHIQUES DE L'IA1
Les préoccupations éthiques sont contemporaines, voire antérieures , à l'invention de l'expression «intelligence artificielle» . En effet, on se rend compte que tous les auteurs qui se sont prononcés sur ce thème ont dès lors insisté sur la nécessité de définir des limites et des principes éthiques pour l'intelligence artificielle .Les préoccupations éthiques se trouvent au cœur même de sa création, de son développement et de son utilisation, mais surtout dans l'hypothèse où des outils artificiels (robots) pourraient être dotés d'une capacité équivalente, voire supérieure, à celle de l'homme, de façon à être capables de prendre des décisions, donc à faire des choix de façon autonome et rationnelle, comme s'ils étaient dotés d'une intelligence et d'une volonté propres.2 Il faut donc bien préciser que l'éthique n'est pas quelque chose d'extérieur, d'externe, qu'on est supposé ajouter à la conception et au développement des exploits de l'IA mais, au contraire, qu'elle doit bien présider à toutes les étapes de son existence, de la même façon qu'elle est toujours imbriquée dans n'importe quelle action humaine à tout moment, y compris les activités et les exploits scientifiques de toute nature.L'étude très importante et tout à fait opportune du CNIL, intitulée«Comment permettre à l'Homme de garder la main? Les enjeux éthiques des algorithmes et de l'intelligence artificielle», publiée en décembre 2017 , l'a très bien signalé, de la même façon qu'elle a souligné les principales étapes du développement de l'IA et des domaines d'application de ses exploits.3 Il convient en effet de bien distinguer tout d'abord les questions éthiques qui se posent dans les phases de création et de développement des algorithmes, qu'on désigne comme étant la source de l'IA, des questions éthiques relatives à son application, et finalement, si tant est que ce soit un jour le cas, au niveau de l'apprentissage automatique(«machine learning») et de la formalisation d'une éthique afin de la programmer dans une machine. Ce dernier aspect sera traité au point suivant.Quant aux étapes,il faut d'abord reconnaître que tout doit commencer par la formation de l'ensemble des intervenants de la «chaîne algorithmique», des concepteurs et techniciens aux professionnels, politiciens et décideurs, en passant par les utilisateurs et les citoyens en général. Il s'agit donc d'une éthique imbriquée dans la culture issue de l'utilisation de nouveaux moyens de communication sociale.Ensuite, l'éthique doit être présente dès la recherche et la conception, jusqu'à la fabrication des systèmes, au niveau de la programmation et de l'apprentissage, de la perception ou de l'image de la machine et du financement des programmes et projets.Quant aux domaines d'application, on devra considérer en principe les suivants comme étant les plus importants :• les transports (planification du trafic, voitures «intelligentes», transports «à la demande», covoiturage et trains sans conducteur);• la santé mobile, les soins à domicile et la robotique médicale;• l'éducation (apprentissage en ligne, professeurs robots, systèmes tutoriels intelligents, traduction automatique);• la politique et la vie communautaire;• la culture et les médias;• la justice;• les finances et l'économie, en particulier les

QUELLE ÉTHIQUE POUR L'IA?1 Toutes les règles juridiques mentionnées devront avoir pour base une définition de l'éthique fondamentale et universelle. C'est-à-dire une éthique qui ne soit pas confessionnelle, mais qui découlera d'un «impératif catégorique qui représenterait une action comme nécessaire pour elle-même, et sans rapport à un autre but, comme nécessaire objectivement» .Ce sera, en effet, dans les principes d'une éthique kantienne qu'on devra aller chercher non seulement les fondements pour toute activité humaine qu'aura pour but l'IA ,mais aussi l'éventuelle conception d'un algorithme spécifique à introduire dans le design même de systèmes algorithmiques d'IA autonomes «pour contrer le caractère de «boîtes noires» que peuvent avoir les algorithmes dès lors qu'ils se présentent comme des systèmes opaques» .2 Selon ces principes, «il y a un impératif qui, sans poser en principe et comme condition quelque autre but à atteindre par une certaine conduite, commande immédiatement cette conduite. Cet impératif est CATÉGORIQUE. Il concerne, non la matière de l'action, ni ce qui doit en résulter, mais la forme et le principe dont elle résulte elle-même; et ce qu'il y a en elle d'essentiellement bon consiste dans l'intention, quelles que soient les conséquences. Cet impératif peut être nommé l'impératif de la MORALITÉ.» Cet impératif catégorique s'énonce comme suit: «Agis uniquement d'après la maxime qui fait que tu peux vouloir en même temps qu'elle devienne une loi universelle» ou bien comme ceci: «Agis comme si la maxime de ton action devait être érigée par ta volonté en LOI UNIVERSELLE DE LA NATURE.» 3 Et KANT précise encore: «L'impératif pratique sera donc celui-ci: Agis de telle sorte que tu traites l'humanité aussi bien dans la personne de tout autre toujours en même temps comme une fin, et jamais simplement comme un moyen.» De ce principe découle que «la moralité consiste donc dans le rapport de toute action à la législation qui seule rend possible un règne des fins. Or cette législation doit se trouver dans tout être raisonnable même, et doit pouvoir émaner de sa volonté, dont voici alors le principe: n'accomplir d'action que d'après une maxime telle qu'elle puisse comporter en outre d'être une loi universelle, telle donc seulement que la volonté puisse se considérer elle-même comme constituant en même temps par sa maxime une législation universelle. Si maintenant les maximes ne sont pas tout d'abord par leur nature nécessairement conformes à ce principe objectif des êtres raisonnables, considérés comme auteurs d'une législation universelle, la nécessité d'agir d'après ce principe s'appelle contrainte pratique, c'est-à-dire, devoir. Dans le règne des fins le devoir ne s'adresse pas au chef, mais bien à chacun des membres, et à tous à la vérité dans la même mesure.» 4 Dans le design même de l'algorithme éthique, il faudra donc prendre en considération que tous ses éléments doivent contenir:« 1.Une forme, qui consiste dans l'universalité, et à cet égard la formule de l'impératif moral est la suivante : il faut que les maximes soient choisies comme si elles devaient avoir la valeur de lois universelles de la nature;2. Une matière, c'est-à-dire une fin, et voici alors ce qu'énonce la formule: l'être raisonnable, étant par sa nature une fin, étant par suite

services financiers (banque et assurances);• la sécurité et la défense;• l'emploi, le recrutement et la gestion des ressources humaines;• la qualité de vie (environnement);• les services à domicile (robots ménagers de la gamme 2030);• les jeux vidéo et d'autres divertissements en 3D.Les principales questions éthiques qui se posent concernent le respect de la vie privée, la protection des droits d'auteur, la responsabilité civile et criminelle, la certification, le marché du travail, la fiscalité et la politique, et surtout l'apprentissage automatique, notamment ce qu'on appelle l'apprentissage profond («deeplearning») et ses implications sociétales. Même si, à court terme, les scientifiques n'envisagent pas la possibilité que les systèmes d'IA puissent choisir de façon autonome de faire du mal aux gens, il sera toujours possible à certaines personnes d'utiliser l'IA de le but de nuire et de porter préjudice.D'où la nécessité croissante d'une réglementation suffisamment contraignante pour prévenir les méfaits de son utilisation.4. Contrairement à ce qu'on dit habituellement, il n'est pas vrai que tous ces domaines ne font pas déjà l'objet de règles juridiques, toutes ces activités ne se déroulant pas dans un vide juridique.Au contraire, dans la plupart des États membres de l'UE, et même au niveau européen, il existe déjà des normes juridiques qui règlent certains aspects comme: les conditions de la collecte et de la conservation des données des personnes ainsi que l'exercice de leurs droits, afin de protéger leur vie privée et leurs libertés; l'interdiction qu'une machine puisse prendre seule, sans intervention humaine, des décisions entraînant des conséquences cruciales pour les personnes; le droit pour les personnes d'obtenir, auprès de celui qui en est responsable, des informations sur la logique et le fonctionnement de l'algorithme . Aux États-Unis, certains aspects concernant par exemple les voitures autoguidées et sans conducteur (règles adoptées par le Nevada Department of MotorVehicles), ou la vie privée , la responsabilité civile et criminelle, les contrats d'agence ou la fiscalité, ont fait déjà l'objet d'une réglementation.Au niveau européen, le règlement général sur la protection des données (RGPD), qui entrera en vigueur en mai 2018, consacre aussi certains principes fondamentaux applicables à l'IA, notamment le fait que l'informatique ne doit jamais porter atteinte ni à l'identité humaine ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques.5 Cependant il est bien vrai qu'il n'y a pas encore de cadre juridique consensuel et harmonisé pour réglementer des aspects fondamentaux de l'IA tels que, par exemple, les fondements pour le développement de l'IA, les principes de l'ingénierie de l'IA et les règles communes pour l'utilisation de l'IA.Dans son étude, que l'on suit de près, non seulement en raison de son excellence quant à l'analyse théorique, mais surtout pour sa méthodologie participative associant la société civile, la CNIL propose deux principes fondateurs: le principe de loyauté , selon lequelil convient d'assurer de bonne foi le service de classement ou de référencement, sans chercher à l'altérer ou à le détourner à des fins étrangères à l'intérêt des utilisateurs; y figurent notamment, d'une part, la pertinence des critères de classement et de

une fin en soi, doit être pour toute maxime une condition qui serve à restreindre toutes les fins simplement relatives et arbitraires; 3. Une détermination complète de toutes les maximes par cette formule, à savoir, que toutes les maximes qui dérivent de notre législation propre doivent concourir à un règne possible des fins comme à un règne de la nature. Le progrès se fait ici en quelque sorte selon les catégories, en allant de l'unité de la forme de la volonté (de son universalité) à la pluralité de la matière (des objets c'est-à-dire des fins), et de là à la totalité ou l'intégralité du système.Mais on fait mieux de procéder toujours, quand il s'agit de porter un jugement moral, selon la stricte méthode, et de prendre pour principe la formule universelle de l'impératif catégorique:Agis selon la maxime qui peut en même temps s'ériger elle-même en loi universelle» 5 Il ne faut cependant jamais oublier – et toujours prendre dûment en considération – que pour KANT, l'autonomie de la volonté est le principe suprême de la moralité. En effet, il considère à juste titre que:«L'autonomie de la volonté est cette propriété qu'a la volonté d'être à elle-même sa loi (indépendamment de toute propriété des objets du vouloir). Le principe de l'autonomie est donc: de toujours choisir de telle sorte que les maximes de notre choix soient comprises en même temps comme lois universelles dans ce même acte de vouloir. Que cette règle pratique soit un impératif, c'est-à-dire que la volonté de tout être raisonnable y soit nécessairement liée comme à une condition, cela ne peut être démontré par la simple analyse des concepts impliqués dans la volonté, car c'est là une proposition synthétique; il faudrait dépasser la connaissance des objets et entrer dans une critique du sujet, c'est-à-dire de la raison pure pratique; en effet, cette proposition synthétique, qui commande apodictiquement, doit pouvoir être connue entièrement a priori (…).Mais que le principe en question de l'autonomie soit l'unique principe de la morale, cela s'explique bien par une simple analyse des concepts de la moralité. Car il se trouve par là que le principe de la moralité doit être un impératif catégorique, et que celui-ci ne commande ni plus ni moins que cette autonomie même» Mais dans cette mesure, on est obligé de supposer la liberté comme propriété de la volonté de tous les êtres raisonnables. En effet, dit KANT:«Ce n'est pas assez d'attribuer, pour quelque raison que ce soit, la liberté à notre volonté, si nous n'avons pas une raison suffisante de l'attribuer aussi telle quelle à tous les êtres raisonnables. Car, puisque la moralité ne nous sert de loi qu'autant que nous sommes des êtres raisonnables, c'est pour tous les êtres raisonnables qu'elle doit également valoir; et comme elle doit être dérivée uniquement de la propriété de la liberté, il faut aussi prouver la liberté comme propriété de la volonté de tous les êtres raisonnables; et il ne suffit pas de la prouver par certaines prétendues expériences de la nature humaine (ce qui d'ailleurs est absolument impossible; il n'y a de possible qu'une preuve a priori); mais il faut la démontrer comme appartenant en général à l'activité d'êtres raisonnables et doués de volonté. Je dis donc: tout être qui ne peut agir autrement que sous l'idée de la liberté est par cela même, au point de vue pratique, réellement libre; c'est-à-dire que toutes les

référencement mis en œuvre au regard de l'objectif de meilleur service rendu à l'utilisateur et, d'autre part, l'information sur les critères de classement et de référencement mis en œuvre; et le principe de vigilance, selon lequel la promotion d'une «obligation de vigilance» viserait à contrebalancer le phénomène de confiance excessive et de déresponsabilisation dont on a vu qu'il était favorisé par le caractère de boîte noire des algorithmes et de l'IA devant avoir une signification collective .Plus importants encore, les principes d'ingénierie doivent constituer les bases de tous les systèmes algorithmiques, dont les principaux seraient: l'exigence d'intelligibilité ou d'explicabilité des algorithmes ; l'introduction d'une obligation de redevabilité ou d'organisation de la responsabilité donnant lieu à une attribution explicite des responsabilités impliquées par son fonctionnement; et finalement, le principe interdisant la prise de décision par une machine seule et impliquant donc toujours la nécessaire intervention humaine, pas nécessairement à l'échelle de chaque décision individuelle mais, par exemple, de loin en loin sur des séries plus ou moins nombreuses de décisions, «collectivisant» en quelque sorte cette obligation, la modulant en fonction de la sensibilité des applications considérées et de la configuration de la balance avantages/risques .Finalement, les principales lignes directrices pour une utilisation correcte de l'IA devraient obéir à des règles communes et uniformément respectées, telles que: former à l'éthique tous les maillons de la «chaîne algorithmique»: concepteurs, professionnels, citoyens; rendre les systèmes algorithmiques compréhensibles en renforçant les droits existants et en organisant la médiation avec les utilisateurs; travailler le design des systèmes algorithmiques au service de la liberté humaine pour contrer l'effet «boîtes noires»; constituer une plateforme communautaire d'audit des algorithmes; encourager la recherche de solutions techniques pour faire de l'UE le leader de l'IA éthique et lancer une grande cause communautaire participative autour d'un projet de recherche d'intérêt général; renforcer la fonction éthique au sein des entreprises .6 Cela dit, l'on pourrait bien sûr être également d'accord avec la plupart des recommandations de la résolution du PE , qui, «considérant qu'il est possible, en fin de compte, qu'à long terme, l'intelligence artificielle surpasse les capacités intellectuelles de l'être humain», juge «qu'il est utile et nécessaire de définir une série de règles, notamment en matière de responsabilité, de transparence, et d'obligation de rendre des comptes, qui reflètent les valeurs humanistes intrinsèquement européennes et universelles qui caractérisent la contribution de l'Europe à la société» et que «ces règles ne doivent pas brider la recherche, le développement et l'innovation dans le domaine de la robotique».7 Dans ce sens, le PE a raison quand il considère que «dans l'hypothèse ou un robot puisse prendre des décisions de manière autonome, les règles habituelles ne suffiraient pas à établir la responsabilité juridique pour dommages causés par un robot, puisqu'elles ne permettraient pas de déterminer quelle est la partie responsable pour le versement des dommages et intérêts ni d'exiger de cette partie qu'elle répare les

lois qui sont inséparablement liées à la liberté valent pour lui exactement de la même façon que si sa volonté eût été aussi reconnue libre en elle-même et par des raisons valables au regard de la philosophie théorique.Et je soutiens qu'à tout être raisonnable, qui a une volonté, nous devons attribuer nécessairement aussi l'idée de la liberté, et qu'il n'y a que sous cette idée qu'il puisse agir. Car dans un tel être nous concevons une raison qui est pratique, c'est-à-dire qui est douée de causalité par rapport à ses objets. Or il est impossible de concevoir une raison qui en pleine conscience recevrait pour ses jugements une direction du dehors; car alors le sujet attribuerait, non pas à sa raison, mais à une impulsion, la détermination de sa faculté de juger. Il faut que la raison se considère elle-même comme l'auteur de ses principes, à l'exclusion de toute influence étrangère; par suite, comme raison pratique ou comme volonté d'un être raisonnable, elle doit se regarder elle-même comme libre; c'est-à-dire que la volonté d'un être raisonnable ne peut être une volonté lui appartenant en propre que sous l'idée de la liberté, et qu'ainsi une telle volonté doit être, au point de vue pratique, attribuée à tous les êtres raisonnables.» 6. C'est en effet sur la base de ces deux idées fondamentales, l'autonomie de la volonté et la liberté, que KANT peut expliquercomment un impératif catégorique est possible.«L'être raisonnable se marque sa place, comme intelligence, dans le monde intelligible, et ce n'est que comme cause efficiente appartenant à ce monde qu'il nomme sa causalité une volonté. D'un autre côté, il a pourtant aussi conscience de lui-même comme d'une partie du monde sensible, où ses actions se trouvent comme de simples manifestations phénoménales de cette causalité; cependant la possibilité de ces actions ne peut être saisie au moyen de cette causalité que nous ne connaissons pas; mais au lieu d'être ainsi expliquées, elles doivent être comprises, en tant que faisant partie du monde sensible, comme déterminées par d'autres phénomènes, à savoir, des désirs et des inclinations. Si donc j'étais membre uniquement du monde intelligible, mes actions seraient parfaitement conformes au principe de l'autonomie et de la volonté pure; si j'étais seulement une partie du monde sensible, elles devraient être supposées entièrement conformes à la loi naturelle des désirs et des inclinations, par suite à l'hétéronomie de la nature. (Dans le premier cas, elles reposeraient sur le principe suprême de la moralité; dans le second cas, sur celui du bonheur.) Mais puisque le monde intelligible contient le fondement du monde sensible, et par suite aussi de ses lois, et qu'ainsi au regard de ma volonté (qui appartient entièrement au monde intelligible) il est un principe immédiat de législation, et puisque aussi c'est de cette manière qu'il doit être conçu, quoique par un autre côté je sois un être appartenant au monde sensible, je n'en devrai pas moins, comme intelligence, me reconnaître soumis à la loi du premier, c'est-à-dire à la raison qui contient cette loi dans l'idée de la liberté, et par là à l'autonomie de la volonté; je devrai conséquemment considérer les lois du monde intelligible comme des impératifs pour moi, et les actions conformes à ce principe comme des devoirs.Et ainsi des impératifs catégoriques

dégâts causés»;il a donc bien fait de demander à la Commission de proposer des définitions communes, au niveau de l'Union, et de veiller notamment à: - créer un système européen général d'immatriculation des robots avancés;- garantir la possibilité d'exercer un contrôle humain à tout moment sur les machines intelligentes;- mettre au point un cadre éthique de référence clair, rigoureux et efficace pour le développement, la conception, la fabrication, l'utilisation et la modification des robots;- doter les robots avancés d'une «boîte noire» contenant les données sur chaque opération réalisée par la machine, y compris les logiques ayant contribué à la prise de décisions;- fonder le cadre éthique de référence sur les principes de bienfaisance, de non-malfaisance, d'autonomie et de justice, sur les principes et valeurs consacrés à l'article 2 du traité sur l'Union européenne et par la Charte des droits fondamentaux de l'Union européenne, tels que la dignité humaine, l'égalité, la justice et l'équité, la non-discrimination, le consentement éclairé, le respect de la vie privée et de la vie familiale et la protection des données, ainsi que sur d'autres principes et valeurs fondateurs du droit de l'Union, tels que la non-stigmatisation, la transparence, l'autonomie, la responsabilité individuelle et la responsabilité sociale, et sur les pratiques et codes de déontologie existants;- créer une agence européenne chargée de la robotique et de l'intelligence artificielle, à même de fournir l'expertise technique, éthique et réglementaire nécessaire pour soutenir les acteurs publics concernés, tant au niveau de l'Union que des États membres, dans leur effort pour garantir une réaction rapide, éthique et éclairée face aux nouveaux enjeux et perspectives, en particulier transfrontaliers, du progrès technique dans le domaine de la robotique.Toutes ces recommandations, cependant, ne nous donnent pas une réponse à la question fondamentale de la nature et de l'essence même de l'éthique, quand on parle d'intelligence artificielle, ni à la question des fondements de cette éthique.

sont possibles pour cette raison que l'idée de la liberté me fait membre d'un monde intelligible. Il en résulte que si je n'étais que cela, toutes mes actions seraient toujours conformes à l'autonomie de la volonté; mais, comme je me vois en même temps membre du monde sensible, il faut dire qu'elles doivent l'être. Ce «devoir» catégorique représente une proposition synthétique a priori, en ce qu'à une volonté affectée par des désirs sensibles s'ajoute encore l'idée de cette même volonté, mais en tant qu'elle appartient au monde intelligible, c'est-à-dire pure et pratique par elle-même, contenant la condition suprême de la première selon la raison; à peu près comme aux intuitions du monde sensible s'ajoutent les concepts de l'entendement, qui par eux-mêmes ne signifient rien que la forme d'une loi en général et par là rendent possibles des propositions synthétiques a priori sur lesquelles repose toute connaissance d'une nature. L'usage pratique que le commun des hommes fait de la raison confirme la justesse de cette déduction. Il n'est personne, même le pire scélérat, pourvu qu'il soit habitué à user par ailleurs de la raison, qui, lorsqu'on lui met sous les yeux des exemples de loyauté dans les desseins, de persévérance dans l'observation de bonnes maximes, de sympathie et d'universelle bienveillance (cela même lié encore à de grands sacrifices d'avantages et de bien-être), ne souhaite de pouvoir, lui aussi, être animé des mêmes sentiments. Il ne peut pas sans doute, uniquement à cause de ses inclinations et de ses penchants, réaliser cet idéal en sa personne; mais avec cela il n'en souhaite pas moins en même temps d'être affranchi de ces inclinations qui lui pèsent à lui-même. Il témoigne donc par là qu'il se transporte en pensée, avec une volonté qui est libre des impulsions de la sensibilité, dans un ordre de choses bien différent de celui que constituent ses désirs dans le champ de la sensibilité; car de ce souhait il ne peut attendre aucune satisfaction de ses désirs, par suite aucun état de contentement pour quelqu'une de ses inclinations réelles ou imaginables (par là, en effet, l'idée même qui lui arrache ce souhait perdrait sa prééminence); il n'en peut attendre qu'une plus grande valeur intrinsèque de sa personne. Or il croit être cette personne meilleure, lorsqu'il se reporte au point de vue d'un membre du monde intelligible, ce à quoi l'astreint malgré lui l'idée de la liberté, c'est-à-dire de l'indépendance à l'égard des causes déterminantes du monde sensible; à ce point de vue, il a conscience d'une bonne volonté qui de son propre aveu constitue la loi pour la volonté mauvaise qu'il a en tant que membre du monde sensible: loi dont il reconnaît l'autorité tout en la violant. Ce qu'il doit moralement, c'est donc ce qu'il veut proprement de toute nécessité comme membre d'un monde intelligible, et cela même n'est conçu par lui comme devoir qu'en tant qu'il se considère en même temps comme membre du monde sensible.»
PRINCIPES ET FONDEMENTS D'UNE ÉTHIQUE POUR L'IA1 Tandis que le progrès et l'évolution scientifiques et technologiques jouent avec les limites du possible et de l'impossible, l'éthique pose la question de la définition des limites du souhaitable et du non souhaitable, même si c'est possible. 2. C'est là que se pose la question de la définition d'une «roboéthique».3. Il n'est pas

nécessaire de remonter au XVIIe siècle et aux discussions entre Descartes et Hobbes, l'un affirmant que la pensée est propre et exclusive à l'Homme et l'autre argumentant que la pensée n'est qu'un calcul mathématique gigantesque, que la pensée humaine serait simplifiable en une formule.4 La question n'est pas de savoir s'il est ou s'il sera possible que des robots ou des systèmes d'IA puissent développer des capacités de raisonnement et de décision égales à l'Homme ou même supérieures en vitesse et en capacité de mémorisation et donc de choix, mais s'il est souhaitable que cela se produise.5 Il est de la plus haute importance, dans ce contexte, que les géants du WEB (Google, Facebook, IBM, Microsoft et Amazon) aient lancé en septembre 2016 un partenariat pour définir de «bonnes pratiques», notamment en matière d'éthique, baptisé «Partnership on Artificial Intelligence to benefit People and Society», qui viserait entre autres à «protéger la vie privée et la sécurité des individus»,«s'opposer au développement et à l'usage de technologies d'IA qui violeraient les conventions internationales ou les droits humains» et à ne «promouvoir que des technologies qui ne font pas de mal» .6 En effet, la question éthique qui repose toujours, chez les Hommes, sur l'autonomie de la volonté ou la liberté de choisir, rendant valide la maxime «video meliora, proboque, deteriora sequitur», même si cette liberté n'est que purement «formelle» dans le sens kantien et non «métaphysique», ne se posera jamais de la même façon chez n'importe quelle IA, puisque, même dans le cas ultime de l'IA dite «générale», capable d'effectuer toutes les tâches intellectuelles qu'un être humain est capable d'exécuter, elle n'aura jamais le dilemme du choix entre le «bien» et le «mal» puisqu'elle sera étrangère à toute notion de valeur.7 On revient donc à la distinction fondamentale entre «personne» et «être intelligent» . Pour n'importe quelle entité intelligente, on doit s'assurer qu'elle se conduise et agisse toujours selon des principes de rationalité stricte. On ne peut pas demander qu'elle agisse en accord avec des valeurs puisqu'elle n'a pas de liberté pour choisir selon des sentiments et l'«autonomie de sa volonté» est strictement contrôlée par le système logique qui est à son origine.8 Bien sûr qu'une option comme celle prévue par certains spécialistes, qui préconisent dès lors la création d'un disjoncteur (kill-switch) ou bouton de réinitialisation (reset-button), permettant de désactiver ou de réinitialiser un système d'IA super intelligent qui s'est emballé, pose toujours la question de savoir si on arrive à temps, et de qui aura le droit/devoir d'actionner ce bouton .9 À mon avis, il faut anticiper toute situation de ce type, et prévoir une sorte d'«éthique by design» introduisant dans toute exploit d'IA un «algorithme éthique» contenant les principes kantiens exposés, de façon à ce qu'il empêche purement et simplement que cette IA puisse violer ces «normes» et donc puisse agir au-delà de ces paramètres.10 Si la possibilité que cette avancée technologique ne semble pas pouvoir se vérifier de ce point de vue et qu'il n'est pas techniquement possible de développer et d'introduire ce genre d'algorithme, alors le principe de précaution exige que tout avancement scientifique soit stoppé précisément à cette frontière.11 Comme

dans d’autres domaines de la science, ce sont les Hommes responsables de la conduite des peuples selon des principes démocratiques internationalement acceptés qui doivent décider des limites infranchissables de certains développements scientifiques ou technologiques qui s’avèrent contraires aux principes éthiques qui ont été décrits et qui seraient en mesure de conduire à la destruction de l’humanité.

Matt	Allison	Vodafone	Vodafone welcomes the opportunity to provide comments to the European Commission's High Level Expert Group AI (AI HLEG) on its Draft Ethics Guidelines for Trustworthy AI. In preparing this work, the Commission has identified three pillars to support and promote the development of ethical AI in the EU: boosting investment, preparing for socio-economic changes and ensuring an appropriate ethical and legal framework to strengthen European values. In our view these pillars correctly address the most urgent policy challenges emerging from the global race to adopt AI technologies. For European policy makers, these challenges are particularly acute, and the way in which we respond will determine the extent to which Europe leads or languishes in the race to become a global AI powerhouse. Our strongest rivals, the US and China, benefit from economies of scale stemming from large domestic markets under a unified regulatory framework. In addition, the balance in both of these markets between unchecked innovation and upholding individual rights is skewed towards the former, placing less restrictions on what AI developers can do within the regulatory framework. Lastly, and perhaps most importantly from the perspective of boosting AI innovation, in both the US and China funding for AI research and development is being made available on an enormous scale. While in the US much of this funding derives from private equity and Venture Capitalist activity, in China the centrally planned economic system has guaranteed significant sums to be made available for AI development under the 'Made in China 2025' economic plan. To compete on a global scale, the EU must be able to leverage its own resources and those of member states national exchequers to ensure that equivalent capital is being invested in indigenous EU AI technology. Correspondingly EU policy makers should attempt to ensure a harmonized regulatory environment across the EU with regards to AI technology to avoid costly fragmentation. We do not envisage that the EU's rights based framework should be a disadvantage to AI innovation, or hamper the EU's ability to compete at the global level. Rather we agree with the European Commission that a strong ethical framework, grounded in fundamental human rights underwritten by the EU's founding treaties, could be leveraged for competitive advantage, by making the EU home to trustworthy AI solutions. Only AI which is trustworthy and robust has the potential to achieve mass market adoption: thus by positioning the EU as the standard bearer of Trustworthy AI from the outset, we are confident that EU technology providers can become the global exporting powerhouses of the future. We need only observe the global reaction to the introduction of the General Data Protection Regulation (GDPR) to see how the exercise of 'soft power' based on a strong rules based framework can indeed help to shape global markets to the EU's competitive advantage. We see no reason why AI ethics should be any different, and as a leading European technology company, Vodafone is strongly committed to working with the European Commission and multistakeholder groups to deliver on this ambition. Our comments on the draft ethics guidelines are intended primarily to ensure consistency with existing	General comments To the best of our knowledge there is no widely accepted international guidance available to determine what is ethical and what is not. Human Rights Conventions and national laws based on these conventions are at present the only commonly agreed articulation of what is ethical. Therefore, ethics should be based on them. Vodafone has conducted an internal mapping of Human Rights Conventions and data processing practices for our internal compliance documentation (not specific to AI). This has been done in context of trying to understand what are the "rights and freedoms of Individuals" that may be impacted by various data processing. We commend this approach (which incidentally has the support of the EUDPS) to the AI HLEG and suggest that this could become a general requirement for companies deploying AI technology.Vodafone supports the adoption of a self-regulatory approach to AI development in the EU grounded upon strong fundamental rights. The introduction of the General Data Protection Regulation demonstrated the ability of EU policy makers to intervene to uphold EU citizen's fundamental rights beyond the EU's borders. The extra territorial effect of the GDPR, combined with the sheer size of the addressable market for online goods and services within the EU has led to a rapid adoption of GDPR like data protection standards across the globe. As such the EU has been able to leverage its scale and soft power to shape the global market for data driven products and services. Ethical AI presents a similar opportunity for the EU to create the global gold standard for secure and trustworthy AI. Vodafone endorses the overarching conceptualisation of the relationship between fundamental rights, principles and values set out at the beginning of this chapter. The cascading relationship from fundamental rights (legal, immutable) down to ethical principles (abstract, high level norms to uphold human centric/trustworthy AI) and finally to values (granular, measureable guidelines for a business to operationalize those rights and principles) is a helpful one.In operationalising these guidelines, providers of AI technology should be empowered to contextualise and make adjustments to suit the various different technology use cases they deploy. Vodafone defines Artificial Intelligence as 'the application of advanced analytical techniques (ML, DL and NLP) combined with automation and related feedback loops to solve problems and seize opportunities in new ways". Our use of AI falls into two broad categories: technology focused AI and commercially focused AI. For the former, AI is being used to assist with fault detection, predictive maintenance, networking planning and optimisation, all of which combines to ensure that we are able to make more efficient use of our physical assets. With regards to the latter, AI is also being used in a commercial setting, through the deployment of Virtual Assistants (more information on our chatbot Tobi can be found here), pricing and promotions, predictive care, smart retail and more. The AI HLEG should establish at the outset that a 'one size fits all' approach is not suitable for AI applications, and that a determination should be made on a case by case basis how ethics guidelines should apply to different AI use cases. Detailed comments AI HLEG text	General comments Mapping the list of requirements for trustworthy AI against Vodafone's internal policy and practices we see a high degree of commonality with the Vodafone approach (our own approach centres on Transparency & Accountability, Ethics & Fairness, Security and Privacy, Humanity and Diversity). The AI HLEG list is somewhat repetitive: it should be made clear in this document if these requirements are ranked in order of importance.An alternative list could perhaps be structured slightly differently and could include some other items: 1. Procedural requirements:o Accountability and Human oversight;o Data Governance;o Impact assessment and Ethics/Fairness by Design (both for new AI technologies and for each new use case of existing technologies, including regular re-evaluation of existing technologies and use-cases)o "Conservative approach" - i.e. if something is tried out for the first time, one should first try it out on a very limited audience, subject to highest level of human oversight. 2. Material requirements (i.e. "criteria" for impact assessments):o Respect of international human rights laws and standards o Legitimacy and compliance with lawso Transparencyo Human agencyo SecurityWe may also want to consider the extent to which some kind of public oversight or governance mechanism is needed (regulatory, judicial or perhaps a parliamentary committee that publishes guidelines). Detailed comments AI HLEG text (Data Governance pp 14): The datasets gathered inevitably contain biases, and one has to be able to prune these away before engaging in training. This may also be done in the training itself by requiring a symmetric behaviour over known issues in the training setVodafone comments: 'Data pruning' is an interesting concept, and one that should be further elaborated on by the AI HLEG. Is there a risk of over deletion/overzealous pruning? We would also appreciate clarification from the AI HLEG on the extent to which this practice differs from "data minimisation" - an established concept under GDPR. Is it perceived as an ongoing process, to check and re-check the data volume/breadth over time? The principle of Human Rights Impact Assessment is critical here. By asking the right set of questions, the potential biases will be more likely to surface. Importantly, while bias is typically driven by data quality, mechanisms can be set up to minimize bias also through the algorithms themselves.AI HLEG text (Data Governance pp 14): Feeding malicious data into the system may change the behaviour of the AI solutions. This is especially important for self-learning systems. It is therefore advisable to always keep a record of the data that is fed to the AI systems, to the extent possible within technical and regulatory constraints (for e.g. GDPR requirements with regard to data erasure). Vodafone comments: Vodafone supports a strong requirement on digital platform providers to ensure appropriate control over the input data being served to AI, to ensure that it is neither malicious, nor in breach of hate speech laws or societal norms. We recognize however that it is very difficult and impractical to monitor real-time data and ask customers for example to use a certain type of language only. Instead of training customers, a reinforced learning system should be trained to either ignore or not use	General commentsVodafone is supportive of the objective of this chapter of the ethics guidelines, to provide a practical checklist of questions to consider in order to ensure the development of trustworthy/human centric AI from an early stage of the development cycle. However, our concern with this section is that while the questions here are appropriate considerations, they lack the technical detail and specificity which would make them a useful or practical tool for AI product development or engineering staff, particularly within a small organization. In general, there is too much repetition of the early sections of the ethics guidelines and not enough effort to provide a clear structure and benchmarking/self-assessment tool for AI developers which is easily understandable in a variety of different languages and AI use cases. Detailed commentsAI HLEG text (Data Governance pp 24): What data governance regulation and legislation are applicable to the AI system? Vodafone comments: We would ask the AI HLEG to consider carefully the linkages which exist between the debate around AI ethics and the ongoing debate around establishing a level regulatory playing field for equivalent services, particularly in the context of the ePrivacy Regulation. Vodafone advocates for a proportional regulatory framework around data to enable AI innovation, with a level regulatory playing field between operators. Vodafone also supports the development of globally harmonised data protection standards, based on free and open data flows to prevent the imposition of unjustified data localisation requirements which would certainly hamper the development of AI. We suggest the AI HLEG could include specific language here on the need for more harmonised cross border data transfers standards i.e. by referring to existing EU standards such as GDPR and the Free Flow of Data regulation. However, Vodafone would strongly reject the introduction of a horizontal "data governance law". Different data is subject to different legal protections arising from a number of different sources. Data governance are the internal practices that ensure those rules are complied with and not something which requires external regulation. Market mechanisms should govern access to data held by private entities, not regulation. From an economic perspective, data is becoming a valuable asset that underpins innovative data-driven business models. Mandatory requirements for data to be made available to public authorities would be a disincentive to invest in technology that would generate data in the future, and would therefore act as a brake on innovation. Operators investing in tools and technology to collect and analyse data should be able to develop a commercial model to for the reuse and aggregation of this data. Sharing should take place on the basis that it is legally valid, socially acceptable and economically viable.AI HLEG text (Governing AI Autonomy pp.24): Is a process foreseen to allow human control, if needed, in each stage? Is a "stop button" foreseen in case of self-learning AI approaches? Within the organisation who is responsible for verifying that AI systems can and will be used in a manner in which they are properly governed and under the ultimate responsibility of human beings?Vodafone comments: This is a critical question. In our view, it is not necessary to	At Vodafone, we are using AI to help to improve our products and services and to run our business as effectively as possible: AI chat bots increase the speed with which we can respond to routine customer enquiries; our 'big data' team uses AI to analyse large, anonymous data sets from customers (who have given us permission) so we develop new and better products and services; and we are deploying AI technology in our mobile networks to identify where capacity is needed so our customers can make calls and access the internet without interruption.As AI grows in usage and impact, we have a responsibility to consider how our use of this technology impacts our customers, our employees, and wider society. We believe it is critical to ensure that the AI algorithms we use are designed to respect both the privacy and security of the data they analyse. We also want to ensure that the insights we derive from big data are fair and not subject to any unintended bias.

regulatory frameworks, and to advise where measures proposed by the AI HLEG are either disproportionate or technically not feasible. In all cases we have suggested alternative wording which should bring the ethics guidelines into line with industry best practice, and ensure a clearer link with AI developments which are currently underway across the telecommunications ecosystem. Detailed comments AI HLEG text (Executive Summary, pp III): Trustworthy AI has two components: (1) it should respect fundamental rights, applicable regulation and core principles and values, ensuring an “ethical purpose” and (2) it should be technically robust and reliable since, even with good intentions, a lack of technological mastery can cause unintentional harm.Vodafone comments: Vodafone considers the concept of 'technological mastery' to be too vague. Suggested alternative: "(2) it should be technically robust and reliable, making use of international cybersecurity standards and best practices to eliminate unintentional vulnerabilities". AI HLEG text (Executive Summary, pp III): In contrast to other documents dealing with ethical AI, the Guidelines hence do not aim to provide yet another list of core values and principles for AI, but rather offer guidance on the concrete implementation and operationalisation thereof into AI systems. Such guidance is provided in three layers of abstraction, from most abstract in Chapter I (fundamental rights, principles and values), to most concrete in Chapter III (assessment list).Vodafone comments: The document describes itself as an operational/practical set of guidelines for developers of AI systems, however it suffers from many of the same problems of other industry standard AI policy/ethical guidelines: i.e. it lacks technical specificity and aspires towards high level principles rather than granular commitments against which companies and individuals can be audited. This may ultimately be the best approach, but we should be careful about the AI HLEG selling this document as something it is not. AI HLEG text (Executive Summary, pp III): In the final version of these Guidelines, a mechanism will be put forward to allow stakeholders to voluntarily endorse them.Vodafone comments: Vodafone would welcome additional detail from the Commission on what the endorsement/certification mechanism will entail in the final draft guidelines. In particular, we would welcome clarity on how this mechanism will dovetail with existing and forthcoming codes and quality schemes at the national, EU and global level. The AI HLEG would be advised to avoid duplication of existing schemes where possible to avoid confusion for consumers of AI solutions and unnecessary complexity for AI developers. AI HLEG text (Executive Summary, pp iv): Moreover, the Guidelines should be seen as a living document that needs to be regularly updated over time to ensure continuous relevance as the technology and our knowledge thereof, evolves. This document should therefore be a starting point for the discussion on “Trustworthy AI made in Europe”.Vodafone comments: Vodafone strongly supports this approach. We believe that to remain relevant this should be a living document, updated over time to reflect new AI insights and market developments. A (Respect for Human Dignity, pp 7): To specify the development or application of AI in line with human dignity, one can further articulate that AI systems are developed in a manner which serves and protects humans’ physical and moral integrity, personal and cultural sense of identity as well as the satisfaction of their essential needs.Vodafone comments: Advocating AI which serves and protects human's moral integrity could perhaps be too onerous for many AI use cases. The concept that neutral technology should be able to 'protect' an individual appears potentially problematic, particularly in scenarios where the AI being deployed has no direct interface with the end user (for example AI technology deployed within an ECS network for the purpose of fault detection and remedy). Suggested rewording: 'one can further articulate that AI systems are developed with full respect to human's physical and moral integrity'. We do acknowledge however that in some circumstances AI may be used to replace human interaction where it may have to undertake a human judgement task. In such cases it is necessary for the agent to preserve and follow established ethical norms for preserving human dignity in any form.AI HLEG text (Citizens Rights, pp 7): At the same time, citizens should enjoy a right to be informed of any automated treatment of their data by government bodies, and systematically be offered to expressly opt out.Vodafone comments: In our view this requirement in the ethics guidelines goes too far. Automated decisions involving personal data are already subject to GDPR requirements. AI HLEG text (Ethical Principles in the Context of AI and Correlating Values, pp 8): in particular situations, tensions may arise between the principles when considered from the point of view of an individual compared with the point of view of society, and vice versa. There is no set way to deal with such trade-offs.Vodafone comments: The distinction between principles considered from the point of view of the individual, compared with that of society is a very interesting and complex one, and warrants greater attention than is granted in this document. Service providers like ours will need to think carefully about how they should act when there is a clear ethical tension between the needs of an individual user and society at large, and should have in place at the very least an ethical code or set of instructions to guide their behavior in these circumstances. It Is our experience that where such tensions exist, it should not just be left to private operators to make a determination on where the correct balance lies between fundamental rights. Such determinations should only be made by actors with a clear public mandate to decide where the appropriate balance should lie; usually judicial authorities or elected officials. Where private operators can play a role is through the introduction of technical measures which reduce the need for difficult tradeoffs: in the context of the ePrivacy Regulation pseudonymisation has been identified as a technical measure which can be deployed to ensure electronic communications data can be used without impinging on fundamental rights. We would encourage the AI HLEG to undertake a detailed study of whether technical measures exist which could reduce tensions between the rights or individuals certain type of data sets, backed up by strong human oversight. AI HLEG text (Non-discrimination pp 16): It is important to acknowledge that AI technology can be employed to identify this inherent bias, and hence to support awareness training on our own inherent bias. Accordingly, it can also assist us in making less biased decisions.Vodafone comments: Vodafone strongly supports the reference to the power of AI to help society identify and eradicate inherent biases and to assist humans in making less biased decisions. The AI HLEG could commit to a more detailed examination of the positive examples of AI being used to tackle inherent bias. Establishing respect for Human Rights as a fundamental parameter of the underlying technology should help produce this result.AI HLEG text Respect for (and enhancement of) Human Dignity: AI products and services, possibly through "extreme" personalisation approaches, may steer individual choice by potentially manipulative "nudging". At the same time, people are increasingly willing and expected to delegate decisions and actions to machines (e.g. recommender systems, search engines, navigation systems, virtual coaches and personal assistants).Vodafone comments: Personalised AI solutions have the potential to vastly improve the consumer experience, saving them money and offering them timely and relevant deals. We are aware however that consumer IoT devices, linked to ubiquitous digital platforms (home assistants) have the potential to influence user behaviour and diminish human autonomy and dignity. Providers of consumer IoT products and online intermediation services should not be able to leverage this technology to unfairly promote or benefit their products and services. In addition, we suggest that the AI HLEG focus on the following key principles: i) transparency as a tool to ensure consumers are empowered to make the right choices and ii) use of AI as a tool itself to empower consumers, a concept which has been further elaborated by ARCEP AI HLEG text (Technical and Non-Technical Methods to achieve Trustworthy AI): An evaluation of the requirements and the methods employed to implement these, as well as reporting and justifying changes to the processes, should occur on an on-going basis. In fact, given that AI systems are continuously evolving and acting in a dynamic environment, achieving Trustworthy AI is a continuous process.Vodafone comments: Vodafone strongly supports the focus on continuous and ongoing assessment of trustworthy AI systems. Trustworthiness is not a static concept, or something which can be guaranteed from one product innovation cycle to the next. AI HLEG text (Ethics and rule-of-law by design pp 19): Whenever an AI system has a significant impact on people’s lives, laypersons should be able to understand the causality of the algorithmic decision-making process and how it is implemented by organisations that deploy the AI system.Vodafone comments: We ask the AI HLEG to clarify on whom should the obligation fall to ensure AI systems are intelligible to laypersons? Service providers should be obliged to be transparent and up front about how AI is being deployed across their networks however we would argue that responsibility must sit with governments and guarantee human control at all levels of AI (e.g. where this is deployed deep in the network for fault detection) Human control may in some cases only be necessary in setting the outcomes, whereas in other cases, where there is a significant impact on individuals, human control is essential. The "conservative approach" and "human oversight" principles should certainly help with this objective. Human control and a stop button failsafe may not be necessary precautions for all types of self-learning AI approaches. Indeed, if we mandate these types of control, even for AI which is deployed deep in our networks and has no human interaction or customer facing element, we could deprive AI of its greatest potential: to solve problems (like cancer treatment, reversing global warming etc.) which humans have not been able to. We request that the AI HLEG give more detailed thought to some of these subsidiary questions before including human control/stop button as a requirement in this section of the guidelines. A contextual understanding of AI, where different use cases are permitted differing levels of human control appears to us to be the optimum outcome.

practical model for this could be to split the guidance into rights/principles and practical guidance levels. The former should remain stable while the latter should be revised in light of technological and market developments at appropriate intervals (every twelve to eighteen months). AI HLEG text (Executive Summary, pp iv): Finally, beyond Europe, these Guidelines also aim to foster reflection and discussion on an ethical framework for AI at global level.Vodafone comments: Vodafone supports international harmonisation with regards to the rules governing ICT and Internet policy wherever possible. We ask whether the AI HLEG may go further than the above statement, in pushing not only for 'reflection and discussion' at the global level, but inviting international partners to adopt the EU model for human-centric/trustworthy AI, either through bilateral partnerships (the recent France-Canada AI declaration) or through multilateral norm and standard setting bodes (ETSI, ISO, GSMA, ENISA).

and society at large as outlined above. AI HLEG text (The Principle of Beneficence: "Do Good" pp 8): AI systems should be designed and developed to improve individual and collective wellbeingVodafone comments: We are fundamentally concerned about the concept of AI systems being employed only for 'good'. 'Individual and collective wellbeing" is not well-defined concept in European Charter of Fundamental Rights, which is the relevant governing legal document here. On a practical level, we believe this could go beyond the remit of private actors, whose main responsibility is towards their shareholders and their customers, as governed by contract. As set out above, use of AI may simply be part of the technical evolution of communications networks and have no "good" or "bad" consequences, in the ethical sense. Vodafone asks the AI HLEG to give more careful thought to these difficult practical and philosophical questions before including a principle of AI beneficence in the final ethics guidelines. AI HLEG text (The Principle of non-maleficence "do no harm" pp 9): By design, AI systems should protect the dignity, integrity, liberty, privacy, safety, and security of human beings in society and at work.Vodafone comments: Concerns regarding AI duty to 'protect' noted herewith (Respect for Human Dignity, pp 7). We refer the AI HLEG to the UN Guiding Principles Reporting Framework which enshrines a duty for states to protect human rights and corporate entities to respect human rights. Suggested amendment: "AI systems should respect the dignity, integrity, liberty, privacy safety and security of human beings"AI HLEG text (The Principle of non-maleficence "do no harm" pp 9): AI systems should not be designed in a way that enhances existing harms or creates new harms for individuals.Vodafone comments: This sentence appears to be missing a vital qualifying concept of intent. Suggested rewording: "AI systems should not be designed in a way which intentionally enhances existing harms or creates new harms". We would also appreciate further detail from the AI HLEG on the specific harms envisaged in this context. AI HLEG text (The Principle of non-maleficence "do no harm" pp 9): To avoid harm, data collected and used for training of AI algorithms must be done in a way that avoids discrimination, manipulation, or negative profiling.Vodafone comments: Vodafone understands this could be quite an onerous requirement given the mainstream practice of profiling users to offer them more targeted services. Suggested rewording: "to avoid harm, data collected and used for training of AI algorithms must be done in a way which which avoids harmful/negative discrimination " and is in line with all applicable data protection and other laws, including the Charter on Fundamental Rights. Again, we are concerned that this principle lacks specificity: If something is intended to be prohibited or restricted, then the harms should be quite clearly articulated. There are unlawful (e.g. racial or gender based) and lawful discrimination (everything which is not prohibited by law - e.g. price discrimination). Is the intention to also limit the lawful discrimination? We ask the AI HLEG to provide additional clarity on lawful/unlawful discrimination in the context of the ethics guidelines. AI HLEG text (The Principle of

public authorities to ensure a higher level of education in relation to how AI works (a basic standard of AI literacy could be a feature of a forward looking educational curriculum). This can only be built up in time just as food or technology literacy has been imbued in the general population. For example, all packaged food are required to have details of their ingredients etc. This may be one of the ways to ensure consumers of AI are similarly kept informed of what is done to their data. In broad terms we would support the proposed policy objective of AI operators having an obligation to explain in understandable terms how the AI in question works (without having to publish the algorithm itself). Also, we again refer to GDPR restrictions on automated decision making. Suggested alternative wording: "…any AI decision making which is likely to have significant impact on the rights and freedoms of individuals…"

Autonomy: “Preserve Human Agency” pp 9): If one is a consumer or user of an AI system this entails a right to decide to be subject to direct or indirect AI decision making, a right to knowledge of direct or indirect interaction with AI systems, a right to opt out and a right of withdrawal.Vodafone comments: The context of the use of AI is very relevant here. If AI is used within a communications network to improve energy efficiency or routing, there should be no need for a consumer to have a right to opt in or out of such use. However, if a consumer’s data is being used to make decisions about that person using AI, they should be informed that this is happening. Where personal data is being used, an individual already has the ability to opt out under the GDPR so there is no need for additional requirements. However, GDPR does not make a distinction between "direct or indirect" interaction - but it refers to "legal effects" produced that affect the individual. This implies the direct or indirect distinction is irrelevant, it's the impact that matters. Vodafone would appreciate clarity from the AI HLEG on the interaction between the principle of autonomy as stipulated here, and the right to opt out of automated decision making as enshrined under GDPR, to assess how the two obligations would interact in practice. What we need is a clear culture or reading the data protection laws in a way that respects Human Rights and their fulfillment in data processing context.AI HLEG text (The Principle of Justice (be fair) pp 10): the positives and negatives resulting from AI should be evenly distributed,Vodafone comments: Vodafone questions the extent to which this is a realistic and practical requirement to include within these ethics guidelines. While all private actors engaged in the development of AI recognize the need to ensure these products and services work for the betterment of society, the requirement for all positives to be 'evenly distributed' is vague. This sort of metaphysical confusion arises out of lack of clearly articulated harms to be avoided, something which we ask the AI HLEG to rectify urgently. In our view most of the metaphysical confusion arises from the fact that the distributions mentioned in the text and which form the basis "Be Fair" principle should be applied to are not specified in advance with sufficient clarity. Grounding AI ethics in a clear legal framework derived from the Charter of Fundamental Rights and pertaining legislation would prevent much of this confusion arising in the first place. AI HLEG text (Identification Without Consent pp 11): Noteworthy examples of a scalable AI identification technology are face recognition or other involuntary methods of identification using biometric data (i.e. lie detection, personality assessment through micro expressions, automatic voice detection).Vodafone comments: Vodafone supports strong consent requirements around the development and use of facial recognition technology. Google's updated AI ethical principles blog explains how they have recently withdrawn a number of AI facial recognition projects on ethical grounds - this is a sensitive area and we recommend a robust approach to upholding individual privacy.. When processing biometric data with the purpose of identifying individuals, GDPR clearly states these are sensitive personal data (Art 9) for which stricter rules

						are needed. We would be obliged in any case to obtain explicit consent for this type of data.		
Anonymous	Anonymous	Anonymous				In my opinion, the section 'Technical methods' might benefit from a more detailed description of the various (fundamental and multi-disciplinary) sub-areas of AI research, as well as related scientific challenges. The following document, which was recently launched by a Special Interest Group representing the AI research community in the Netherlands, might be useful input for this (see page 3-5 in particular): http://ii.tudelft.nl/bnvki/wp-content/uploads/2018/09/Dutch-AI-Manifesto.pdf		

Leena	Sarvaranta	VTT Technical Research Centre of Finland	As said in the report, the principles, and principals as such, do not have enough power to carry us far enough in the discussion of AI ethics. Concrete advice is needed by developers of how to act and what kind of factors to take into account. Indeed, presenting "check lists" or "heuristics" should be avoided, as ethics is always negotiable. The starting point for the report is good: The quality of life of people and "good life" should and can be the only justification of technology. Good life can be mirrored with the help of values that individuals or communities follow. Here the main question pertains to the interpretation of 'good'. What can be considered good, from whose perspective, and what kinds of choices generate an increase in goodness? Analysis of people's ways of life and life settings, should be starting point to technology design, involving stakeholders in discussion, and using relevant data in design processes.We must take serious consideration of warnings about technological imperative, the situation where the development of technology has autonomy, i.e., its own logic, in which individual actions or hopes are of secondary value. Consequently, a justification crisis can emerge, as much of the AI technology can be in want of legitimating values. If the justification of AI technology is not presented clearly enough, the ultimate goals of its use will become unclear. The threat of misusing AI and ending up in ethical crises in its use can be addressed by promoting constant discussion of the very essence that justifies and guides human action and the use of AI. As dealing with ethics is always contextual, it is also negotiable. The discussion should be carried with various stakeholders, using effective governance activities. In this sense, the report could mention the concept of RRI (Responsible Research and Innovation) which is broadly used in European research. Scope of the Guidelinespage 3: …"It is, therefore, explicitly acknowledged that a tailored approach is needed given AI's context-specificity.". Could it be described somewhere in the Guidelines concretely how this is envisioned to be implemented (Chapter 2?) as it now keeps a bit hanging in the air.B. A Framework for Trustworthy AIPage 3 (Ethical Purpose): …"this section can be coined as governing the "ethical purpose" of developers, deployers and users of AI..." How can the users be governed by the Guidelines, be obliged to behave ethically? What is the role of schools for example?	The Principle of Non-maleficence "Do not Harm" page 9: … "AI systems should not harm human beings...". How is the Principle applied in Defence context? Section 5.4 (LAWS Lethal Autonomous Weapon Systems, page 12) touches on this issue, but limited to the context of autonomous systems; this conflict also exists for non-autonomous settings (e.g. a simple, non-autonomous, image recognition AI can be used for harmful purposes).Page 9: …"AI algorithms must be done in a way that avoids…, or negative profiling….". It is questionable if such a regulation / governance is workable that allows profiling, but only positive profiling!page 9: …"Vulnerable demographics (e.g. children, minorities, disabled persons, elderly persons, or immigrants) should receive greater attention to the prevention of harm, given their unique status in society…." Is there a danger that this introduces bias if it is done in an unbalanced manner?5.2 Covert AI systems page 11: ... "A human always has to know if she/he is interacting with a human being or a machine…". This is right, and yet, this is somewhat situational as well. It has turned out in automatic news production that news readers do not especially wish to know if e.g. the sport news article is written by a human journalist or a machine. The important thing is that the facts are right in the article.5.5. Potential longer-term concernspage 12: As mentioned, AI implementations are currently still done mainly by well-trained scientists. However, that may change in the further future, and "everybody" will be able to develop and run an AI application at the push of a button. This will require education (from early age onwards) about ethics and do's and don'ts. As stated, even if probability of certain risky long term developments is very low, yet the potential harm could be very high. Using the classical formula 'risk probability x potential harm' motivates to keep the developments under observation.	2. Data Governancepage 15: …"in large enough data sets these(misjudgements, errors and mistakes) will be diluted since correct actions usually overrun the errors"… However, it should be pointed out that you cannot always rely on assumption that big numbers take care of errors!4. Governance of AI Autonomy (Human Oversight)page 16: …"This also includes the predicament that a user of an AI system, particularly in a work or decision-making environment, is allowed to deviate from a path or decision chosen or recommended by the AI system"… Why is the term 'predicament' (=unpleasant situation) used here? After all, this is a totally acceptable and usual situation. Especially in the higher level decision making in e.g. healthcare, the end decision is always made by a human, who may take into account the AI's recommendation, or not. We are talking there about decision support instead of decision making by AI.8. Robustnesspage 17: …"Hacking is an important case of intentional harm"… This definition is a bit narrow. Hacking is not always done to cause harm, it can also be done for the good - e.g. to test and improve systems (ethical hacking).2. Technical and Non-Technical Methods to Achieve Trustworthy AIpage 18: Is it necessary to stress the distinction between technical and non-technical methods? HLEG AI Guidelines can have a great impact in improving future development work by stressing the importance of multidisciplinary design and strong interaction between developers from engineering and human & social sciences backgrounds. Ethical issues concerning adoption and use of technology-supported services are raised and solved in a social, political and economic context, and in the context of use. Ethical issues related to the introduction, adoption and use of AI technology should always be contextualized. How the ethical dilemmas are solved depends on the attitudes and views of the different formal and non-formal stakeholder groups involved. The risk for technological imperative will not be decreased as long as technology development is carried out in a silo, and as long as non-technical development is discussed separately. Figure 3 (page 19): The report does not discuss how other requirement on AI systems – like those coming from the market/consumer, society companies, etc will play out against the requirements that spring from the the Rights, Principles and Values to the left in Figure 3. Should these other requirement be totally overrun by the ethical one? In practice, it will not go like that.1. Technical methodsEthics & Rule of Law by Design (X-by-design)page 19: …"Compliance with law as well as with ethical values can be implemented, at least to certain extent, into the design o the AI system itself…" NB:Although human rights are permanent statements, ethical values are significant and lasting ideals, shared by the members of a community, about what is good or bad and desirable or undesirable. Thus interpretation of human rights may vary depending on the context? Ethical assessment of design decisions calls thus for conscious reflection of ethical values and choices in respect to the context the technology is intended to be used. Thus, it is not possible to derive values from facts ("Hume's guillotine") and leave it to a machine to do? How can good life be formulated in formal (technological)	1. Accountabilitypage 24: Continuity of relevant staff would be good to have. If data scientists (either at university or at company) come and go on short temporary contracts it may become very difficult to have continued, guaranteed skills and knowledge. Assessment List for Healthcare DiagnoseOverall, the proposed list in the Draft Guidelines fits well. Perhaps some issues can be weighted more than other for the different use cases, but the general contents are there. For healthcare specifically, we should concentrate on:- Accuracy - often we don't know what is the right output (no Gold standard, inter-expert variation in diagnosis, exact diagnosis only available after death etc). How to deal with that?- Knowing when not to give an output (e.g. if the input is too far from the training data)- Issues with adaptive systems that change their behaviour. How to guarantee performance over time?- Explainability/Transparency- Thorough validation and testingIn addition to the Assessment List presented in the Draft Guidelines, it might be worth asking the following questions:• What kind of influence does the solution have on the users' quality of life?• Does the solution enhance the quality of life of the users better than any other artefact or solution?• What needs (and whose expectations) should the solution fulfil?• Who benefits from the solution? Would other stakeholders benefit from it?• What are the possible alternatives for solving the problem?• How should the users (direct and indirect) be seen, interpreted and understood in the design?• How are the users involved in the design theoretically and empirically?• Is the main basis for the design answering the users' needs and expectations?• What are the multiplicative effects of the solution?References: Saariluoma, P., Cañas, J.J., & Leikas, J. (2016). Designing for Life - A human perspective on technology development. London: Palgrave MacMillan. ISBN 978-1-137-53046-2 DOI 10.1057/978-1-137-53047-9Leikas, J., & Kulju, M. (2018). Ethical consideration of home monitoring technology: A qualitative focus group study. Gerontechnology 2018;17(1):38-47; https://doi.org/10.4017/gt.2018.17.1.004.00.Leikas, J., Koivisto, R., & Gotcheva, N. (2018). Ethics of autonomous systems. In Heikkilä, E. (Ed.) Effective autonomous systems. VTT Framework for developing effective autonomous systems. VTT White Paper, December 2018.	VTT supports this draft document which is highly relevant and much needed! Contributions of VTT's comments were given by (name.surname@vtt.fi): Mark van Gils, Jaana Leikas, Caj Sodergard, Lula Rosso and Leena Sarvaranta(1) The sections in the documents should be numbered hierarchically. (Now e.g. on page 19 it reads "1. Technical methods. The numbering should be "III.2.1. Technical methods)(2) Page iv, GLOSSARY: Definition of AI is better here than e.g. that in Wikipedia. However, a couple of comments:" Artificial intelligence (AI) refers to systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving (NB below) their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take (according to pre-defined parameters) to achieve the given goal. AI systems can also be designed to learn to adapt their behaviour by analysing how the environment is affected by their previous actions".- "…perceiving their environment…". Many applications, that claim they are AI, do not sense any environment, which is more specific for robotics. E.g., software for analyzing medical images like melanoma . The application gets the image data maybe from a database and makes conclusions on it. It does not require any new imaging to be done for completing information. Robotic Process Automation (so called software robots) are another service, that is difficult to fit into the definition above.3. Page iv, GLOSSARY: Definition of 'Bias' l here is just one possible interpretation of bias (prejudice against something or somebody). 'Bias' can also relate e.g. to a too positive estimation of AI performance. Also, in the context of AI research, the word 'bias' has a specific meaning in the architecture of neural networks, which is very different than what is described here.4. Page iv, GLOSSARY : 'technically robust and reliable' could be added as term in the Glossary (just as 'ethical purpose' now is.

language if the parameters cannot be (and should not be) based on unambiguous facts? Using abstract principles as a basis for technology ethics requires discussion and constant updating of views (discursive ethics), and is strongly linked with governance and RRI (Responsible Research and Innovation). In this discourse, the voice of citizens should actively be heard. This discussion cannot be left for the AI to carry out. Otherwise, we are facing technological imperative, where the development of technology has autonomy and its own logic, in which individual actions or hopes are of secondary value. Then, AI could end up being in the role of a dictator, who defines values for good life. In a case like this, is there a risk that we are setting aside traditional values as legitimating entities for our actions? Will AI create new values to follow?The sentiment of the possibility to use AI for self-assessment of ethics of technology is partly reflected in the current version of the document. However, as said already, values are discussed common agreements, and cannot be defined in formal language. They should be discussed and agreed by people. Thus, the document should stress this importance of "non-technical", soft methods where people are used in defining the ethics parameters for AI. Testing & Validatingpage 20: The section on "Testing and Validating" is important and it could be mentioned explicitly that this is an often a very costly and time-consuming exercise that requires rigour and knowledge, and is in many cases woefully underestimated. There is space and need for education here (for AI developers as well as eg research funding agencies). For healthcare applications this is especially a big issue. Moreover, Testing it is important to note that not only the AI system itself must be tested, but also the overall IT system, of which it is a part.Explanation (XAI research)page 21: An additional subheading could be added, "Confidence indication" (or it could be an extension of "Explanation". An AI should be able to tell how sure it is of its output (confidence intervals, error margins or similar), also it should indicate if it is not able to make a reliable decision for the given input and is "just guessing" (e.g. if the input data is completely different than what was in the training set).2. Non-technical methods The Draft Guidelines seem to cover a rather restricted set of non-technical methods for achieving ethical AI. Most important in AI development, along with technical investigations, is to guarantee that understanding of humans and human life is involved in the design work, from the very beginning and in every phase of the design and development work. The Draft Guidelines address this issue by referring to necessity of "Diversity and inclusive design teams" Education and awareness to foster an ethical mindsetpage 22: This is perhaps not AI specific, but it could be added that education in how to correctly interpret e.g. performance measures of AI-based systems (accuracy, sensitivity, specificity, precision, recall etc) and how to translate them to risks and actionable insights in real-life (where e.g. prevalences of different classes may change) is highly needed. This is important especially in health applications, but the understanding of these type of issues (both by AI developers and e.g. healthcare professionals) is often inadequate.

					Developers have wrong ideas about how what the performance of their system is in actual life, and doctors mis-interpret the associated risk that e.g. a medical test output gives. This causes harm on many fronts. This has perhaps more to do with insufficient offering of relevant education in statistics than AI, but it does have a negative effect on dealing correctly with AI too.Diversity and inclusive design teamspage 22: …"diverse in terms of…professional background and skillset". Diversity here is not enough. Professionals with social- and human science -background are needed in the actual design teams to guarantee conceptual and contextual understanding of the AI use case at hand. Conceptual investigations and philosophically informed analyses include questions such as how values are supported or diminished by particular technological designs, who is affected, who benefits from technology, and how should competing values (e.g., access vs. privacy, or security vs. trust) be considered in the design, implementation, and use of the given system or application. Contextual investigations involve social-scientific research on the understandings, contexts and experiences of the target user groups of the technological designs. They focus on the human response to the technology, and on the social context in which the technology is situated.		
Hadrien	Valembois	INTA - International Trademark Association	No specific comments on the introduction.	On the Principle of Justice - Be Fair (page 10): Human review of AI decisions should be included as part of the Principle of Justice. While Governance of AI Autonomy (Human oversight) is mentioned later in the document, INTA stresses the importance of human intervention when legal consequences are involved. The Principle of Justice demands the possibility that automatized decisions with legal impact may be revised by a human being. Human review should be done by individuals who have not participated in the programming of the corresponding AI system, to prevent possible bias.	On the Data Governance Requirement to Achieve Trustworthy AI (page 14): INTA would like to add that AI solutions must rely on transparent and precise data, especially when used by governmental entities to make decisions. For example, in the trademark field, if a government trademark agency uses an AI software program to decide if a trademark can be registered or not, but that program has a database which is incomplete or inaccurate, the decision it will reach would be inadequate. INTA believes that Adequate Data Governance must rely on transparent and complete data. Regarding Standardization (page 21): INTA acknowledges the importance of accreditation systems, professional codes of ethics and standards for fundamental rights compliant design. We recognize the fundamental role of trademarks and intellectual property rights to secure trustworthy AI systems.	Regarding the Use Case of "4) Profiling and Law Enforcement" (page 28): INTA stresses the importance of human review, transparency and completeness of the data, and explicability. INTA understands that it is essential that those who use AI to make decisions are able to explain which was the process to reach that decision, what were the parameters loaded in the system, and what data and algorithms were used.	After careful analysis of the guidelines, due to the generic nature of the document and its intention to provide high level principles across all industries (not directed specifically to the practice of law or trademarks), INTA considers that specific comments on the ethical dimensions of intellectual property (IP) or intellectual property rights (IPRs), since they are not part of the scope of the draft guidelines, would not be applicable at this stage.
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential

Beatriz	Bonete	everis Spain (Chazz Design Studio)			- To integrate an exploratory phase of research that is not only focused on design, but as a way of gathering information about potential applications of AI from an ethical approach. That would mean using a combination of techniques:o Qualitative techniques: to reach a deep understanding and get meaningful insights from a human approach.o Quantitative techniques: to validate statistically and be well-prepared to make decisions before designing AI.- To integrate an evaluative research phase: as a systematic exercise to learn from the applications and uses of AI from people who are using them. It would also be recommended to have experts on Ethics and social issues (diversity, gender perspective…) in this phase to transform the insights and learning into recommendations for future implementations of AI designs.		I´m a sociologist and specialized in research methods and, now, the Head of Research in Chazz Design Studio. Having a specific stream of research in our studio means that in Chazz we are not only "testing" the products/services we design. Testing is important to see if something is going to work, but you also need insights and context to understand the potential and current uses of a product/service, and for that, you will need thorough research. In the draft they say that "traditional testing is not enough." And it´s not. That´s why I would like to focus my contribution in the area of research methodology, as one specific area within Part II: Realising Trusthworthy AI, point 2 "Technical and Non-technical Methods to achieve Trustworthy AI.
Huma	SHAH	Coventry University	The Introduction could include some text that to ensure Trustworthy AI, AI companies, developers and educators need to find ways to democratise the field so that it encourages more females into considering it as a worthy career, engages citizen scientists to investigate ethical usage and interact with communities more so that hidden gems in socio-economically disadvantaged homes are included to appreciate benefits, understand risks of not being involved. Trustworthy AI should not exclude consideration of ways to bridge the Digital Divide. There is a mention in Chapter 1, on page 9, and Chapter 2 page 22, however, a pointer in the Introduction regarding AI being more inclusive so that development of more trustworthy AI tools and technologies are derived that respect the fundamental rights of human beings, and don't lead to barren regions with humans out of work and little prospects for life improvement. Education and inclusivity are also key here to produce human-centric AI for the majority of humans.	Respect for human dignity: should it be noted here, the application of CRISPR gene-editing has already produced twin girls, in November 2018, where their "embryos altered to make them resistant to HIV", as reported here: https://edition.cnn.com/2018/11/28/health/genetic-editing-he-speaks-int/index.html If this technology advances with better machine learning applied to genes' investigation, will humans born through this sort of technology enjoy lower health insurance premiums, for example, and be favoured against humans born the 'normal way', who might then suffer discrimination because they are considered less healthy? In the section on 'The Principle of Non maleficence: "Do no Harm", third line of the first paragraph, is 'freedom of identity' actually meant where 'freedoms of identify' is stated? In this same section, where data is mentioned, should some text on adherence to GDPR is expected from AI services? In the section on 'Principle of Justice "Be Fair"' it might be useful to cite the Angwin et al. (2016) article published in 'Pro Publica' on Machine Bias: software applied in the US justice system was found to be biased against black suspects: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing In the section on 'The Principle of Explicability: "Operate transparently"' should it be made clear that developers of apps, especially smart 'phone apps targeted for children should make transparent what trackers are embedded, so that parents know what usage is being tracked in the apps, then they can pursue the developers to ask what is being done with the tracked information? This is to respect the child's right, preserve their dignity and their autonomy. In section 5.2 'Covert AI Systems', could it also mention the consideration of stereotyping machines/robots/humanoids, when is there a need for a female embodied robot? It might be useful to cite anthropologist Professor Jennifer Robertson's work in this area, including 'Gendering robots: Robosexism in Japan', please see here:	Section 3 Design for all – again, including citizen scientists and engaging more with the community could ensure AI technologies are developed for wide usage. Innovative thinking is needed by AI companies, and educational establishments collaborating here. Section 7 Respect for Privacy – could mention Developers should think carefully why they need to embed trackers when developing any AI technology. Marketing should not be the 'be all end all' position on new tech. Section 8. Robustness: for data to be accurate it needs to be diverse and broad to prevent errors. Here you could cite the case of New Zealand's online passport system that rejected an application, because it failed to recognise the eyes of the applicant, of Asian origin, were open: https://www.reuters.com/article/us-newzealand-passport-error/new-zealand-passport-robot-tells-applicant-of-asian-descent-to-open-eyes-idUSKBN13W0RL	Chapter 3 'Assessing Trustworthy AI' appears compact with valid questions for organisations, and educational establishments training AI specialists.	Thank you for the opportunity to read and offer contribution. Please note I am a supporter of the C.L.A.I.R.E. network. This report is a welcome, timely and necessary document to leverage discussion ensuring development of Trustworthy AI systems, especially in the wake of the Facebook/Cambridge Analytica data harvest scandal leading to the UK Information Commissioner's office bringing a case in court last week against the parent company of the latter: ICO vs SCL in UK Hendon Magistrates court, January 9, 2019 resulting in SCL receiving a fine of £15,000 for failing to "comply with an enforcement notice": https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2019/01/scl-elections-prosecuted-for-failing-to-comply-with-enforcement-notice/ AI is multi-disciplinary not just about learning, it is related to 'what to do' with the knowledge gained. Please recall Alan Turing's seminal 1950 paper, 'Computing machinery and intelligence' was published in a philosophy journal, Mind, for wider readership. To develop trustworthy AI it needs embedding of, for example, psycho-linguistics and socio-linguistics in conversational AI/natural language/dialogue systems (Amazon's Alexa and similar home assistants), as well as to understand historical context, anthropological, philosophical, economics, material science - for lighter, smarter materials for exoskeletons to mobilise the invalid, and robots in manufacturing to mitigate potential damage and injury, as well as mathematical, bio-chemical (cyborgs) and physics. Thank you for time reading this feedback.

https://lsa.umich.edu/histart/people/faculty/jennyrob.html

Section 5.5 ‘Potential longer-term concerns’ might include the risk of 'opportunity to access' AI technologies could further increase the Digital Divide. So the ‘harm’ could be further obstacles to social mobility and increasing family and community well-being and economic prosperity.

Another long-term concern is the competitiveness from nations such as China, Russia and India. So they need to be involved in discussions on developing trustworthy-AI, to ensure EU citizens are not adversely affected by using AI tech developed in non-EU countries. Please see these articles:
CBS 60 minutes: China and AI: https://www.cbsnews.com/news/60-minutes-ai-facial-and-emotional-recognition-how-one-man-is-advancing-artificial-intelligence/?ftag=CNM-00-10aab7d&linkId=62315284

Are China, Russia winning the AI arms race? https://www.reuters.com/article/us-apps-ai-commentary/commentary-are-china-russia-winning-the-ai-arms-race-idUSKCN1P91NM?utm_medium=Social&utm_source=twitter

Microsoft to set up 10 AI labs, train 5 lakh youth in India //economictimes.indiatimes.com/articleshow/67555285.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst

With respect to artificial consciousness, we must also note that humans are suspicious of others, as can be seen in the way some think and talk about ‘others’ (hear how some in the UK talk in the Brexit debate). Some humans will also act negatively towards intelligent machines, exemplified in the 2004 ‘I, Robot’ movie where the human policeman chases a robot running with a handbag wrongly assuming it is a thief without thinking, why would a robot need a human handbag anyway?! Robots would be designed smartly to hold important items within their casing leaving their limbs free to work and cooperate!

Nicolas	Beaume	BrainInSilico (self-employed consultant in bioinformatics)	First I would like to thank you for this draft document and your work on this important topic.I think the introduction overview quite well the question and defined some core values (trustworthly AI, human-centred AI, ethics, type of solutions). I missed only two elements :- the use of AI in academic research- how our european model will compare to other AI strategies in term of ethics and developpement.I am not mentioning the use of AI in academic research only because I am a researcher, but because AI weight more and more in scientific results and this provide an "hidden" influence of AI to knowledge and thus decision. I think than mentioning academic uses of IA in the introduction may be a reminder that before being applied for business, adminstration or well-being, AI is (and will) influence the way we see the world, which is quite important.Other AI policies are not mentioned through the document and I think that it could be interesting to see how the EU position may differs/converge from/with other big AI actors.	I have no special input for part 1-45.1 (identification without consent)The point raised for automatic identification in a mass (such as in security application, face recognition in public places to prevent crime for instance) is right. As a citizen, I would say I would be more open to the use of such data for the police and other official security forces and absolutely not for private companies. Of course, issues can also be raised with official security forces. Identifying possible criminal intentions through face recognition is not as intrusive as doing the same from social networks (which could be easily used for less ethical means such as political identification). Here, the answer is most probably non-technical: restraining the number of persons that have access to the data, controlling them as well.5.3 was not very clear for me. I think example may be useful here5.4Will all expect peace to remain forever and in this context, it is difficult to discuss about LAWS. As a citizen, I won't be chocked that EU develop such weapon to keep the Union able to defend itself should the worst happen. As an alternative, won't it be possible to devlop anti-LAWS system to make ennemy's LAWS inefficient and thus avoid to devlopp such system our self ?5.5I think the possible future of AI can be divided into two parts:1/ AI access to self-consciousness2/ AI overwhelming humans in (almost) everything.1/ is probably more distant than 2/. AI acess to self-consciousness will trigger discussion about what is humanity and sentience. I think, at the end, we don't want to repeat the mistakes done while discovering other cultures and we may want to accept AI as other sentient being and work with them.Thus, preparing this event from know is probably a good idea.It would be even more important if 2/ already happen or is about to. If AI become more efficient than human, and remain a tool, we will have to deal with the fact that human will be less (no more ?) needed in term of employment and society. If AI becomes conscientious and more efficient than humans the question of dominance of AI over humans will raise, for them and for us. We most probably don't want a world dominated by AI, but if we are less efficient than them, our best chance is to have something to offer, a place in this world. So better to think about it before.I don't think it is much too early to start to think about that. At worst we would have too much time, at best, we will not be surprised if overefficient and conscious AI raise before expected.	1-2 data governanceI think is issue is especially important as, in general, data policies, including in academic research tend to be too weak.In the academic, decision maker and project leaders are not the same that AI user. Most of the time they don't anticipate the possible use of AI approaches and thus lack to take into account data policies into the project design.Data are gathered (usually at high cost in biomedical research, genomics for instance) but not designed for AI use, which may lead to bias or issue in assessing the system true performances.As more and more scientific/R&D projects are funded through EU, adding a criteria of evaluation of the data policy to ensure project leaders have taken this point into account would be a good practice.1-4 GovernanceFor some cases, medical AI for instance, the emphasis must be put on the "decision-helper" aspect of the AI. Enforcing this idea on practitioners and the public would greatly help to create trust on AI in field with high impact such as medicine.1-5 biasHere again, enforcing technical solutions such as looking at the learning curve, would greatly help AI-users to detect possible bias and citizen to trust that there are solutions to the bias. 1-8 robustenessas machine learning developper, I would be interested by a UE-recommandation with a concrete list of methods, maybe associated with a high-level publication from UE AI scientists2-1 explanationThis goal is especially hard to achieve in science where we sometimes ask an AI to find hidden correlation we, human not able to think above 3-4 dimensions, are not able to see. The day we are able to make an efficient AI to high dimension problem (such as genome-level prediction of diseases) which explains its solution, EU will become the top leader of science (and many scientists will have to find another job...)2-1 standardizationI think one element which absolutly requiered standard in science is data production. In genomics, most small to medium lab cann not afford to produce enough data to feed an AI but if true standards of sequencing, for instance, would exist, multiple lab could combine their effort to produce enough data or such lab could take advantage of open data produce by other european projects.On the other hand, a standard shall not become so heavy that it slow down innovation...2-1 educationThis is a very big issue, especially for citizens. AI required a bit of technical to be understood enough to trust it. From my experience, even among highly educated people (from non-AI fields) this is a huge effort they are most often not ready to make. I think enrolling AI specialists into vulgarization will be essential to create trust of citizens	For healthcare diagnose and treatments :1 accountabilityit is really a trade-off between how deep the AI can go in analysis (and thus influencing the final decision) and how much the practitioner is responsible if things goes wrong. Also, medicine remains a realm of exceptions and it is especially difficult to be sure that the patient will response positively to the treatment. Most probably, keeping the same level of accountability than now is a good short-term solution2 data governanceAlready discuss before. for scientific/R&D project funded by the EU, enforcing a data policy et evaluation before the project start would avoid many bias in the results/project failure. Project leaders must be aware of the importance of data in AI analysis and take it into account, not only worrying about the cost of data generation !4 gouverning AI autonomyThe risk of scientific AI to become an unwanted sentient AI is low. On the other hand, the complexity of the question and the importance of the quality of the results means AI must stay as much as possible "decision helper". Even if we later demonstrate that AI outperform humans in diagnose tasks, keeping a human overseeing the results and transfer them to the patient is highly desirable.8 robustnessprediction efficiency of AI in diagnoses are very divers. For some tasks, the AI perform almost without error but usually the underlying biological mechanisms are well known. Sometimes, it has good performances in genral but some special cases are still beyond its reach. In any cases the performances are often precisely known but in the second case the consequences could be dramatic. I think that, beyond robustness, it is important to evaluate the type of error made by the AI in diagnoses and treatment assignation to be able to detect these errors and to know when to trust AI and when to ask for deeper investigation. Technical solutions such as AI combination (majority vote approach or having a special sub-AI only in charge to detect special case) could solve this issue	I think that I would say again how I appreciate this work and the fact to open it to stakeholder. I hope my comments will be helpful to the AI HLEG.I have three concluding remarks1/ taking into account the AI impact in academic research seems to be an important thing, as the way scientific AI are use may shape our understanding of the world, in a good or a bad way2/ Discussing with my surrounding, including scientists (but which are not AI expert) shows that citizen tend to understand the risk of AI but are unaware of the solutions that prevent these risks. A huge effort toward citizens sensibilization to the AI field is highly desirable.3/ comparing our EU view of the AI with other AI-powers such as the US, China, Russia, Canada, India or other seems important in the final document.As european citizen, I totally acknowledge with the idea of trusthworthy AI and this is exactly what I want for Europe. Other AI-powers may be less regarding in term of AI conception and ethics and will get a short term advantage in AI leadership, unless we put much more effort than them. It is fine if at the end we get better AI that them, but if it is an accepted strategy, then it should be explained to our fellow citizens who may not understand why China can cure cancer thanks to AI while we are still building ours and to the private sector who may want to push hard to sacrifice regulation for competitivity. I think if we stick to trustworthy AI, we may have to prepare a plan to deal with the gap compare to other powers. We may also think on how to avoid to be "invaded" by non-EU AI that don't meet our standard but may be cheaper and more quickly available.All the best for continuing this important work et vive l'Europe !

Allon	Bar	EXECUTIVE SUMMARYp.I:- "AI is key for addressing many of the grand challenges facing the world"... --> There are huge expectations in this statement, which may or not come to fruition. A simple 'could be' instead of 'is' would avoid criticism of technological utopianism.- The guidelines are "addressed to all relevant stakeholders developing, deploying or using AI". Perhaps this should be expanded (here and elsewhere in the document) to people being *affected by* AI, so that they can use these guidelines to hold the developers, deployers and users of AI to account.- The Executive Guidance prescribes to "Provide, in a clear and proactive manner, information to stakeholders (customers, employees, etc.) about the AI system's capabilities and limitations, allowing them to set realistic expectations." --> Critically, the stakeholders who should be informed are people affected by the AI system's operations, as well users of the system.GLOSSARY- p.IV: Ethical purpose. These two words are repeatedly and prominently used in this document, but I find them confusing and weakening as to what this document should aim for. 'Purpose' usually refers to either a motivation (reasoning) or a goal (determination). While they are relevant in some way, they are not what ultimately decides the impact of AI systems on human beings. I would argue that *process* and *outcome* are more critical to that. Or perhaps more broadly, it may be helpful to think of an ethics life cycle, which can include phases of design, development, implementation, initial use, further use and closure after use. These phases may sometimes be sequential, and sometimes simultaneous.A. RATIONALE AND FORESIGHT OF THE GUIDELINES- p.1: "This working document articulates a framework for Trustworthy AI that requires ethical purpose and technical robustness. Those two components are critical to enable responsible competitiveness, as it will generate user trust and, hence, facilitate AI's uptake." --> There are two large assumptions in here that I don't believe hold true. One, is that ethical use and technical robustness will generate user trust in AI systems; whereas it's A) not even clear if people know when AI systems are being used and B) individuals' trust may be more dependent on whether the system are actually effective in doing what they were intended to do. The second large assumption is that if there is increased user trust more people will start using AI. This may be true in a business-to-business setting, where business users can, theoretically, make a choice whether they want to use system X or (non-AI) system Y. But in the case of the many individuals being *affected by* AI decisions, they rarely if at all have a say in whether they want to be subjected to an AI system (or more broadly, automated decision making). AI systems are being deployed without individuals consenting to be subjected to automated rather than human-based decision making. So more 'trust' on their part will not affect AI's uptake. Lastly, AI's uptake should not be a goal in itself: AI's uptake should only be promoted if and where it improves societal well-being (which may or may not be the case, depending on the system, use case and context).- p.1: similar to the above, "Without AI being demonstrably worthy of trust, subversive consequences may ensue	- p.5: The third § suggests that fundamental rights require an account of the ethical principles to be protected; whereas I see it reversely (and it's also phrased that way on p.6). Overall, while 'rights' and 'fundamental rights' are referenced throughout this document, it's not necessarily clear why an ethics framing is chosen rather than a human rights framing. It might be useful to provide an explanation. Human rights have universal values, have legal and moral power, and have been applied to the technology sector as well. (See Cath & Van Veen: https://points.datasociety.net/artificial-intelligence-whats-human-rights-got-to-do-with-it-4622ec1566d5). Ethics is both more vague as well as lacks the enforcement capabilities that the human rights framework is offering. - p.5: "Building on the basis of decades of consensual application of fundamental rights in the EU provides clarity, readability and prospectivity for *users, investors and innovators*." --> Here again different stakeholder groups are mentioned. Sometimes it's "developers, deployers and users", sometimes the above, sometimes something else. Please check for consistency across the document.- p.5: "Informed consent requires that individuals are given enough information to make an educated decision as to whether or not they will develop, use, or invest in an AI system at experimental or commercial stages" --> As mentioned for the introductory section, it is also about informed consent about whether individuals agree to be subjected to decisions of an AI-based system, something which is often imposed on them. This is different from using a system, they are often not users, but are being affected by the system.- p.7: While it may make sense to group rights here, it seems like a glaring omission to not explicitly mention the right to privacy. It is (one of) the first to be negatively affected by the deployment of AI systems.- p.9: It is unclear how AI helps to "increase citizen's mental autonomy". Also, it's unclear what "trust optimisation towards users" means.- p.9: "possibility to refuse AI services", this and the very top of page 10 seem to be the rare nods to this critical part of autonomy and choice, which should be more prominently featured in the rest of the document.- p.10: "Transparency is key to building and maintaining citizen's trust in the developers of AI systems and AI systems themselves" --> I would add "implementers, organizations running" to the list of actors to be trusted.- p.11: "Differentiating between the identification of an individual vs. the tracing and tracking of an individual, and between targeted surveillance and mass surveillance, will be crucial for the achievement of Trustworthy AI." --> This makes it sound like, unlike tracking/tracing/mass surveillance, identification of an individual and targeted surveillance is acceptable, which I wouldn't so blankly agree with.	- p.14: "developing, deploying and using AI" --> these three aspects do not cover the full breadth of the AI system life cycle, see my comment for p. IV.- p.14: 1. Accountability. --> This is indeed an important aspect, but it seems the text here does not talk about accountability as a whole, but rather about 'remedy' or 'grievance' mechanisms, which is just one aspect of accountability. Accountability is also necessary when things don't go wrong; to help prevent wrongdoing; for people to be able to know what policies exist, to verify their execution, and to address outcomes with whoever's responsible. Accountability thus also includes governance (assigning and exercising responsibility), transparency (covered under the 10th requirement), legal responsibility and indeed also ways to address errors, wrongdoing, etc. (which can be achieved by remedy mechanisms). Chapter III (Assessing) does correctly look at accountability in broader terms.- p.15: 4. Governance of AI Autonomy (human oversight). --> I had a hard time understanding the text in this section. If the focus is on "ensuring appropriate levels of human control" then I would focus the text on that, explaining that depending on the application and on how critical the impact of the automated decision making is, a different level of human oversight or even interference is necessary .- p.17; 7. Respect for Privacy. --> Here, as previously mentioned, the roles of 'user' and 'individual about which the AI system makes a decision', are conflated. Moreover, the text refers to someone's "interactions with the AI system" that needs to have privacy and data protection, but it does not address data used by the AI system that is collected about individuals by other means, for example from third parties or from these individuals' interactions with non-AI systems. In general this section should address the re-use and combining of data from various sources, as well as when data can be discarded and what happens with the data when the AI system is no longer in use.- p.17: 8. Robustness. This sections refers to "secure" algorithms, the meaning of which eludes me. Under 'resilience to attack', hacking is portrayed as an 'important case of intentional harm', whereas hacking is also used for (intentional) security protection, i.e. preventive actions to investigate weaknesses in systems.- p.18: 9. Safety. The description of this section seems to widely overlap with the previous (Robustness) section.- p.18: 10. Transparency. I would explicitly add guidance here about openness about used training data, how it is obtained, and what biases it may contain.- p.19: Here the 3 phases are referred to as "design, development and use", which is inconsistent with the three phases mentioned in my comment for p. 14.- p.19: Ethics & rule of law by design: this section is confusing. Yes, it make sense to incorporate those principles, but it's not clear what the "ethical and legal" rules are that should be complied with. This can then hardly be called a 'technical method'. This section rightly calls for companies to "to identify from the very beginning the ethical impact that an AI system can have", and this is the part that requires more prominence in this document. Meaning, organizations taking in active part in AI systems should investigate the impacts of these systems before putting them to use.	- It's great that there's an assessment list. But the assessment should critically also focus on *impact*. Human Rights Impact Assessments have become more widely adopted and may provide helpful guidance.- p.24: Second paragraph: "[...] i.e. that will have an impact on decision-making processes of individuals or groups of individuals." --> I would leave out "decision-making processes of", as it's the impact on individuals that matters. AI technologies oftentimes take away individuals' ability to take decisions. That automation is part of the benefits AI should provide, which as noted can come with reduced human autonomy.- p.26: Privacy. I would add these bullet points: "How can individuals view, change and control the information about themselves that the AI system uses?", and "How is data minimization adhered to in relation to AI systems' (re-)use of data?"- p.26: Autonomy. I would add this bullet point: "Do users keep their autonomy to ultimately decide based on their individual preferences and circumstances?"- p.28: Key Guidance, bullet 2: "Trustworthy AI is [...] about a continuous process of identifying requirements, evaluating solutions and ensuring improved outcomes throughout the entire lifecycle of the AI system." --> Rather than speaking of requirements, solutions and outcomes, here I would focus on "exploring, identifying and mitigating potential negative impacts, while supporting positive impacts"	Thank you all for all your work in putting together this document. And thanks for giving the opportunity to provide feedback. The above is a selection of key concerns I have regarding the document. When going into more detail I have a few more comments regarding phrasing and technical details, but given the amount of comments that I had on the substance I have left things out. The document contains many good parts, which I haven't commented on as they did not require amendments. The other comments are intended as constructive feedback to help improving the final outcome of the document.- Overall the document is a helpful contribution to the field. But it also feels uneven in quite a few respects. For example, the focus on 'technical robustness' seems to be more mindful of quality and security, the importance of which I do not wish to neglect, but that I don't think should be this document's primary concern. Anyone wishing to successfully deploy AI should and will focus on those aspects in any case. There is a risk that organisations using AI will refer to the latter as an easy win ("look, we are ethical as we are focusing on security"), without taking the impact on individuals into account, which is what I believe this framework should focus on.- The document says it strives to give guidance on implementing and operationalizing principles elsewhere established, but since these are rarely made clear or explicit, it remains rather abstract what people should be doing. See also my comment regarding page 5 (fundamental / human rights).- While the text generally refers to fundamental rights, principles and values that should be respected, it sometimes, also adds regulatory compliance to this mix. This is inconsistent, or at least it's not clear why regulation sometimes is and sometimes is not relevant.- There is some branding (and marketing) of 'Trustworthy AI' sprinkled throughout the document, which may be less relevant. For example, on page 23 it says that stakeholders should be trained in Trustworthy AI. Would this be more important than training in human rights, and training in security? - The document mentions in a few places that it aims to be human-centric. Nevertheless, it is not clear to me if any human research has been conducted in its drafting. In other words, how to know if something is human-centric without having actually asked individuals? This is partly a pedantic point as I realize it's not likely such an effort will be made in this drafting process. It should nevertheless be part of the assessment. - This brings me to something that I find missing generally in this document: using people-focused design approaches in developing and designing AI systems. Research (usually referred to as 'user research' in the technology sector) can help to surface how individuals are, affected by, use, understand, and would like to use AI systems. Any human-centric system aiming to be ethical should prioritise these aspects, especially to make sure that the system serves human needs in the first place. The guidelines rightly point out the need to take into account minority groups, people with accessibility challenges, and others who may be vulnerable to the systems' impact. Thorough research may help unearth specific needs or circumstances for how a diverse of people (who may, e.g., be different from the systems' developers or

			and its uptake by citizens and consumers might be hindered" --> again there seems to be an assumption here that individuals will somehow make a 'choice'. Then what is this choice? How can people 'trust' something that is forced upon them? Should the choice whether or not to be subjected to an AI system's decisions-making itself (as to some extent is done in the GDPR) be promoted as a critical goal in this document? - p.2: Trust in AI includes... I believe this should also include trust in the people or institutions building or deploying them.- p.2: Purpose and target audience. Governments are mentioned as regulators and potential developers/deployers/users of AI, but should also be mentioned as the guardians of the wellbeing of their citizens, and thus having an interest in implementation of these guidelines.	The third section (Assessing Trustworthy AI) provides useful guidance, and should be referenced here too. In addition, one could say that companies should conduct human rights impact assessments. By and large, these are non-technical methods.Lastly, the end of this section refers to security principles, it's not clear why they are added to 'ethics' and 'rule of law' by design. - p.20: Traceability & Auditability. "Evaluation by internal and external auditors can contribute to the laymen's acceptance of the technology." --> Why is "laymen's acceptance" a goal? Here and elsewhere in the document, there seems to be an inverse cause-effect relationship. While possibly unintended, this kind of phrasing makes it appear as if AI technologies MUST be used, and whatever resistance individuals ("laymen") have needs to be simply overcome. Once they've accepted their fate (of having to use this technology), all will be good.Personally, the goals that I believe should be pursued are, 1) to help individuals assess whether technology can be trusted, and; 2) to make technology more worthy of their trust. If these goals are met then acceptance will follow on its own accord.- p.20: Traceability & Auditability. In the last sentence, it is unclear to me 'who' needs to undergo a digital transformation.- p.21: Regulation. The second paragraph here seems to be exclusively focused on remedy mechanisms, rather than regulation. Is there a header missing?- p.22: Codes of Conduct. The ambition that organisations sign up or endorse the Guidelines can be worked out further. How would they do this? To what end? Would there be any verification of the implementation of these Guidelines by organisations that adopted them?		designers) can or would use a system, or be affected by it. The same can be said for gender, as developers of AI system may or may not reflect gender distribution in the real world or the gender distribution of people affected by the AI system. (The document rightly calls attention to this.)- In various places the document refers to 'citizens' or 'European citizens', which leaves out people living in Europe who are not EU citizens, people living outside of Europe affected by AI systems deployed in Europe, as well as people who are affected by AI systems exported outside of the EU (and thus possibly outside of its legal protection). In general, I would recommend referring to 'individuals' or 'people', instead of 'citizens' as more inclusive language.- Along the same lines, there sometimes seems to be an assumption that 'users' are the same as individuals who are being affected by an AI system. While on occasion that may be the case (as for example individuals use a banking app that uses AI to detect malicious entry into the app), but often it is not (for example as an insurance agent uses a software system with AI to decide whether an individual qualifies for health insurance; the former is the 'user' of the system, whereas the latter is the individual affected). It is therefore important to make sure that the language in this document refers to not only users but also to individuals affected by AI systems.
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Jan	Kleijssen	Council of Europe	• Statements without scientific references: The introduction seems to take for granted the superiority of the benefits of AI development over the risks without any documented scientific reference or demonstration to support this statement. The rationale could have been limited to the fact that there is no broad scientific consensus on the concept (difficulties in reaching agreement between experts in certain points of the document itself) and that to face uncertainties, it was necessary to accompany developments with ethics. The reference to "personalised medicine or more efficient delivery of healthcare services" (p.1) is too optimistic and is not sufficiently balanced: improvements are expected, but in very specific tasks and medical professions knows the benefits and limits of statistics.• Too wide scope and limited focus: The attempt to tackle in an abstract and general way all the various developments qualified as "AI" fails to define the specific issues to certain applications. For example, industrial robots do not raise the same issues as the decision support of doctors or judges. In more operational developments, it would have been appropriate to categorise the applicable principles according to different processing families asking specific questions. By pretending to be	In general:• Reference only to EU Treaty and EU Charter of Fundamental Rights: Since reference to human rights is frequent in the document (or "human-centric"), the main international human rights instruments should be referred to as well, in particular theEuropean Convention on Human Rights but also, in this context, Convention 108 (data protection) and the Budapest Convention on Cybercrime. • Lack of reference to the precautionary principle: Given the uncertainty still hanging over the discipline, a translation of the precautionary principle in the field of AI might have been relevant (see for instance Paragraph 2 of article 191 of the Lisbon Treaty about environment). The development of commercial, industrial or public sector applications should be limited to what is known to have the most added values and the least risk (and this fully contributes to trust building). Risk assessment studies should be mandatory.• Lack of reference to gender: Given the figures on the gender gap regarding women as students and professionals in the ICT sector (about 16% in the EU), this aspect should be strengthened in the guidelines in order to a) fulfil commitments in relation to gender equality, women's empowerment and equally shared economic growth and b) because	• Mix of fundamental rights, technical action and policies without any classification: The list of the 10 conditions for the implementation of a Trustworthy AI has been drawn up in alphabetical order, assuming that they are equivalent in importance. This choice, which is understandable for the reasons set out in the document (in particular the need to take into account their specific context of application), may not sufficiently enlighten the reader: fundamental rights (having retained their title or with an effort of reformulation) are thus mixed with technical actions or policies of action. • In the state of the document, some groupings may be considered:- Accountability (requirement 1) could be linked to transparency and the ability to explicability (requirement 10)- The issue of anonymisation (or pseudonymisation) of data in data governance (requirement 2) would be more appropriate in terms of Privacy (requirement 7)- The title "Design for all" (requirement 3) could have remained linked to the notion of equality. The content here is closely linked to the issue of non-discrimination (requirement 5). The question of the digital gap should also be considered separately here by distinguishing between those who cannot and those who do not want to use a computer (to access to a	No further specific comments will be made on the list of questions to analyse the conformity of a development with the stated principles as they are linked to the previous chapters.	

comprehensive and value-based, by limiting the focus on some rights, by considering that "AI" is good (see above) and by the objective set (convince the public to trust and use amply AI Systems), the guidelines could even be a bit counter-productive (the label could be granted to applications which do not respect Human rights).• Lack of references on Human rights: From a legal point of view, it is to be regretted that some rights have been selected at the expense of others or that they have been grouped together for the sole purpose of this study. In general, the limitation to the existing legal framework of the European Union is regrettable (see further comments in Chapter I). • Lack of reference to concrete positive use of AI to strengthen Human rights: Detecting bias, discrimination (gender, etc) could be a goal in itself for policy makers and built fairer society. Being "human-centric" is not enough to produce that result.• Lack in definition of AI: On a technical level, the substantial effort to define AI (an entire document is devoted to it in addition to the glossary) fails to recall the foundations of the technologies used and developed in recent decades: mathematics, statistics and probability. Developments therefore do not appear to take sufficient account of the (extremely rich) academic achievements relating to the limits of these formalisms in general (e. g. C.S. Calude, G. Longo, The Deluge of Spurious Correlations in Big Data, 2017) or to accurately model certain phenomena (e. g. social phenomena, see P. Jensen, Why does society not allow equation?, Seuil, 2018 or R. Nuzzo, Statistical Error, Nature, 13 Feb. 2014). Machine learning algorithms are, as is well known, essentially at the root of the digital industry's renewed interest in advanced task automation since 2010. Their place and complexity deserve to be further developed and clarified, insisting in particular on the fact that each category of machine learning (supervised, unsupervised, reinforcement) has very specific case of applications, that needs different safety precautions (this kind of statement should be enough, without entering in too much details).

ensuring an increased presence of women would also be a way to reduce gender bias. The over-representation of men in the design of these technologies could undermine decades of advances in gender equality but also deepen labour market and power inequalities, given the (ever-growing) importance of this sector. The recent report on artificial intelligence commissioned by the French government focuses on this aspect and for example proposes a target of 40% women students in the IT area.More specifically:• Selected and consolidated Human rights without convincing explanations (B.I.3 p.7): It can be noted that the discussion in is limited to selected fundamental rights, without convincing explanation of the choice that are made. In any case, it must be recalled that all Human rights are inalienable and must be protected. The consolidation of select Human rights for the purpose of that document seems partial, the examples given to illustrate seems sometimes too narrow. It would have been logical to use its 6 pillars of EU CFR (Dignity / Freedom / Equality / Solidarity / Solidarity / Citizenship / Justice) for the document than to invent a new categorisation.More specifically, about principle 3.3 (Respect for democracy, justice and the rule of law): The issue here is not only to deal with interference in elections (where we should add elections influenced with corrupt intent) or to respect the law (this is not an option). The issue that could have been addressed is the replacement of the rule of law by algorithmic calculations in certain situations (allocation/limitation of rights). Laws are legitimised through democratic processes and cannot be slowly replaced by regulation operated by algorithmic systems, the design, operation and control of which are in the hands of very few people. • No link between Fundamental rights and Ethical principles + correlating value (B.I.4 p.8): The work of the AI4People project (inspired by the bioethical approach) was directly introduced here. In consequence, the links announced p.6 (Figure 2) between Fundamental rights, ethical principles and values seem quite artificial. However relevant the work of AI4People may be, the thoughts on principles and values could have been complemented by other ethical frameworks (quoted in the document, such as EGE Statement or even extended to others such as The Toronto declaration). • The "do no harm" principle should be stronger / more specific regarding gender equality aspects due to the high risk of big data reproducing stereotypes leading to potential discrimination / disadvantages for women. It is important to stress that women are not a group but represent half of the population and that they are also over-represented in groups that are considered vulnerable in this context, namely the elderly and persons with disabilities. Therefore a truly ethical purpose and "people-centred" policy in this area can be effective only if gender equality aspects are taken into account.• Critical uses of AI appear anachronistic in a first abstract and theoretical chapter (B.I.5 p11): Even if it is mentioned that there is no clear consensus among the group of experts, this point, referring to concrete applications, appears anachronistic in a first abstract and theoretical chapter. It might have been less controversial to identify here the scientific criticisms that were the subject of a public service).- The developments on the Governance of AI Autonomy (requirement 4) are relevant but abstract - the degree of autonomy of a system cannot be set out of the context of its field of operation (industrial machine, autonomous vehicle, assistance in decision-making in medical or judicial matters, etc). The imperatives to be addressed are quite different depending on whether the mechanism produces decision-making that has consequences for humans or not.- The developments relating to machine learning in the section on non-discrimination (requirement 5) are relevant and could perhaps have been included as soon as the document was introduced and/or in the document on the definition of AI.- The manipulative effects described in the section on "respect of Human Autonomy" (requirement 6) would have been relevant in Chapter I (B.I.5 p11).- Technical robustness is key (requirement 8) but this point would have been an opportunity to contextualise in 3 steps (before development of an AI system, during development and after development) the different actions to be carried out. About the resilience to attack, a specific requirement on integrity of a system could be added. It could be emphasized that systems/data can become corrupted but also conceived/created in a corrupted manner (to harm, commit crimes etc). A general principle could be stated like that: "Systems and/or data must not be or become corrupted".- The Safety requirement (requirement 9) would have been an opportunity to introduce the precautionary principle (key also in legal liability mechanisms as duty of care), with the need to carry out risk studies. The document could also have taken on the responsibility of encouraging research for those areas where uncertainties remain while limiting / prohibiting commercial applications.- Transparency (requirement 10) is a controversial qualification for which the term explicability could have been substituted (deep learning transparency is technically impossible).• Lack of content about the importance of the choice of ML algorithm (B.II.2.1, p19): There is a lack of development on the choice of models and algorithms among the different available (supervised or unsupervised learning, by reinforcement), which is essential to obtain reliable results according to the scope of application (e. g. S. Raschka, Model Evaluation, Model Selection, Model Selection, and Algorithm Selection in Machine Learning, Nov.2018). This is an essential and fundamental recommendation.• Extend the analysis (B.II.2.2, p21): Closer coordination between international organisations, while respecting their mutual prerogatives, should be encouraged in order to strengthen efforts to achieve a common definition of phenomena and coordinate regulatory efforts.There is also a lack on how to build digital services (especially public services) in a world where only 47% of the population has an access to internet: policy makers should be aware to drive change by taking into account that some citizen could not have (or do not want to have) an access to digital services.

				consensus and from which means of action could have been deduced in the following sections. Thus, this part would also have benefited from thoughts on the limits of mathematics and statistics to deal with certain phenomena in a general way (see comments above under Introduction: Rationale and Foresight of the Guidelines). Structural effects on society such as those described in Chapter II, requirement 6 (respect of Human Autonomy) would also have been relevant here.				
Aki	Cheung	Privacy Commissioner for Personal Data, Hong Kong, China	AI is not an end in itself, but a means to achieve the greater end, i.e. improve the well-being of the human race. Therefore, we agree with the draft Guidelines that AI has to be human-centric and respectful of fundamental human rights. These are the two necessary attributes of a Trustworthy AI. The draft Guidelines provide helpful reminder and foundation to organisations in their development and building of AI guide.	The draft Guidelines propose that a Trustworthy AI has to have an ethical purpose. It goes on to elaborate the five fundamental rights of human beings and the five ethical principles of AI for ensuring the ethical purpose. The fundamental rights and the ethical principles are comprehensive and commendable. However, most (if not all) of the fundamental rights identified in the draft Guidelines are not absolute. Sometimes it is justifiable to sacrifice individuals' rights to achieve a wider public interest. An example is the use of AI and facial recognition technique to apprehend terrorists. Hence, the Guidelines may consider allowing certain derogation from the fundamental human rights in exceptional circumstances if the derogation can be justified by necessity and proportionality. The draft Guidelines also seek public feedback on a few contentious issues, and they are discussed in below: • Paragraph 5.1 – "Identification without consent" raised a concern about the use of individuals' consent as a basis for processing of personal data. We suggest that consent should not be heavily relied on because of the risk of "consent fatigue" and the fact that consent alone does not provide additional protection to the rights and interests of individuals, except for respecting the individuals' autonomy. Therefore, where a data processing technology is not clearly allowable by existing laws, the organisation concerned should be required to demonstrate that the legitimate interest of the technology overrides the fundamental rights of individuals and are in line with the ethical principles. Obtaining individuals' consent alone may well not suffice in the circumstance. • Paragraph 5.2 – "Covert AI systems" suggested that AI developers and deployers	The draft Guidelines has identified 10 requirements for Trustworthy AI. All the requirements are commendable. It may consider adding the 11th requirement, i.e. continued attention and vigilance about the risk of AI. This is one of the two founding principles of AI as suggested in the French data protection authority's (i.e. CNIL's) report on AI issued in May 2018. This is a response to the unpredictability of AI. This also echoes with other parts of the draft Guidelines (e.g., Technical and Non-technical Methods to achieve Trustworthy AI) which reiterate that evaluation and assessment of AI systems should occur on an on-going basis.	The assessment questions suggested in the draft Guidelines are comprehensive. That said, the following questions may be added to provide clearer guidance: • Regarding Part 3 – Design for all: o Who will be potentially impacted by the use of the AI system, and how? • Regarding Part 6 – Respect for privacy o What are the sources, nature and sensitivity of the data used in the AI system? o Are there any legal or contractual restrictions on the use of the data? o What is the legal basis for collecting and processing the data under the relevant data protection laws? It is suggested that the assessment questions should be categorised into different stages in the entire lifecycle of the AI system so that a user of the Guidelines will be clear about what matters to consider under each stage of the data life cycle.	The draft Guidelines are well-thought-out, insightful and comprehensive. However, there seems to be a lack of integration between Part I (Respecting Fundamental Rights, Principles and Values – Ethical Purpose) and Parts II & III (Requirements of Trustworthy AI & Assessing Trustworthy AI) of the draft Guidelines. An integration of Part I (the rights, values and principles) into Part III (the assessments) will make the Guidelines wieldier and more likely to guide organisations to achieve the ethical purpose.

				should ensure that human are made aware of – or able to request and validate the fact that – they interact with an AI identity. We agree with this proposition. This is a manifestation of respect to human dignity. More importantly, it reduces the risk of manipulation of people's thinking by using of AI. • Paragraph 5.3 – "Normative & Mass Citizen Scoring" suggested that citizens should be allowed to opt out from scoring. We agree with this proposition because it is a respect to citizens' autonomy. However, the more important issues around citizen scoring are transparency of the scoring systems and the rights of citizens to dispute the scores, especially when the scores have significant impact on the citizens.			
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Stefania	Multari	Confartigianato Imprese					Confartigianato Imprese's comments on EU working document "ETHICS GUIDELINES FOR TRUSTWORTHY AI"Apart from the highly technical definition of AI contained in the defining document, we do share almost all the concerns and solutions raised in the ethical guidelines, and especially the concept of trustworthiness as the center of any discourse on developing AI in the future. It is likewise highly shareable what the report states in the concluding paragraph about Europe as the hotbed of a human centered approach to the future that must be consistently pursued also in the case of AI.The only remark that we consider important to make is about the highly feared, and often misunderstood, impact of AI on the economic system and labor market, especially concerning the substitution of even medium skilled workers with AI based solutions. This perspective raises highly understandable concerns among public opinion throughout the world, and calls for forward looking solution combining innovation and new policies in many fields, including regulation of the labor market, training and educational policies, a new welfare system.As stakeholders, we demand that such concerns will consistently be taken into account in developing future guidelines for AI in EU economic area. This does not mean hampering the development of innovation in the Continent, but stating that such innovation as well as EU's competitiveness cannot be reached at the expenses of its fundamental yet potentially weak elements such as SMEs and workers, that might risk to be marginalized by a borderless technological development.In concrete, we ask to add a new point at issue 10 of the Assessment list (Tranparency) regarding the assessment of social consequences of AI based solution development in terms of labor, to be measured in terms of a 2 years balance between jobs loss and new jobs creation.
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential

Patrick	Haggard	University College London			I strongly endorse the draft guidelines' emphasis on explainability, and my comment is confined to this point. For AIs to be ethical, we need to understand, in principle, HOW they work. I would add additional emphasis on WHY we require explainability. Explainability is part of improvability, which is part of a general societal commitment to progress and greater flourishing. This commitment is ethically important. When a technology system makes an error, it is important to understand the source of the error, and protect people from similar errors in the future by adjusting the system as necessary. The inquiries that follow rail and air accidents are a good example of this working in practice. There is no such mechanism yet in many AIs based on deep neural networks. When the system makes errors, the suggested response is often to enlarge the training dataset. There is no clear guarantee that this will prevent the error from happening again. An important part of the ethics of new technologies is that there be a clear and convincing pathway towards better outcomes, and this involves being able to explain, fix and learn from errors. AIs which do not afford this ability raise potentially severe ethical concerns.		
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	
Luís	PEREIRA	Universidade Nova de Lisboa	1- No explicit emphasis is placed on the AI creation of wealth and its actual distribution among all humans. AI will actually ever more strongly accentuate the increasing wealth gap, unless new social compacts are put in place, there being dangerous risks of resentment and revolt otherwise, and ensuing shunning of AI, a pity because it is after all a conquest of humanity as a whole. The whole question of societal wealth and values is being given short shrift or swiped under the rug. 2- Machines, whether robots or software and their combination, will themselves have to act morally to be convivial with us (and amongst themselves). But we know too little about our own ethics and how to impart it to machines. More ethics research is required, starting now. 3- Similarly, more jurisprudential conceptual scaffolding is needed that will support laws, regulations and standards, including the use of LAWS and autonomous machines in general. 4- The Guidelines should foresee regulations and monitoring concerning the activity of contract consortia, such that individual responsibility is clearly defined from the start -- the so-called "Problem of Many Hands." 5- Joint EU initiatives such as CLAIRE, and international collaboration centres (viz. CERN), should be spelled out as natural venues for increased and widespread value of AI, at the same time striving to avoid the most pernicious dangerous aspects of an AI race, by joint validation, certification, monitoring, and agreed joint AI security. 6- International rules of commitment should be fostered, subscribed and monitored, like	1- The issue of societal values concerning wealth distribution is skimmed over in this chapter. AI will increasingly and acutely widen the pre-existing and wealth gap already on the increase. Not enough concern is shown in the Guidelines regarding the unstoppable encroaching of machines into the heretofore human monopoly of cognition and hand-eye coordination, and overall negative impact on unemployment. The immense technical progress brought about by AI is not being accompanied by a concomitant social progress that will benefit everyone's actual wealth and less striving for a living, not just for the owners of patrimony and technology. 2- The old capital/labour split needs urgent revision. After all, my body is my own limited capital, so even after I leave a company for another, the body capital I spent in the first should continue to benefit me thereafter if that company is successful.	1- Computer languages need to be developed that enable the specification, validation and monitoring of ethical constraints in programs. 2- Programmed AI machines must be subject to safety and compliance tests before being marketed. A case in point are driverless cars, which must comply with common standards imposed by authorities, who thereby become jointly responsible for untoward incidents as a result of improper certification. 3- A recent law that went into effect in California already in 2019, prohibits software that impersonates a human. That should be easy to rapidly obtain consensus on. 4- Large windfall profits should commit to a margin to help promote trustworthy AI by independent organisations.	1- International chartered bodies are needed to enact and assess the trustworthiness of AI and be enabled to denounce violations. 2- Independent and credited auditors must be set up, over and above internal auditing by companies, governments, and protected individual denouncing of risks.	1- Stakeholders must include the Humanities, since the impact of AI is quite wide and needs contributions from a diversity of fields of knowledge, that must be promoted to best contribute. Specifically, I point out Philosophy, Psychology, Ethics, Jurisprudence, Linguistics, Anthropology, Sociology, Economics, Political Sciences, Evolutionary Science. 2- AI research, largely construed, should be further concentrated, centred and promoted in the universities (and research institutes), and there it can easier and more naturally be interdisciplinary in character. 3- A tax on sales is needed, over and above that on profits (always hard to audit because of globalisation and fiscal paradises). 4- A tax on robots and software fully replacing humans must be contemplated, for replacing means replacing, including social security contributions by the worker and the employer. That will help prevent social disruptions.

			with climate change agreements.				
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Wernher Marco	SCHORLEMMER	IIIA-CSIC, supporter of CLAIRE		In addition to the rights, principles, values and requirements for Trustworthy AI discussed in this draft ethics guidelines, we would like to draw attention to the development, deployment and use of AI as is occurs in the context of current power structures, because they may perpetuate and even worsen situations of dominance, control, exploitation and inequality. An important European principle is that of 'subsidiarity', the principle that social and political issues should be dealt with at the most immediate or local level, and that a central authority should perform only those tasks which cannot be performed at a more local level. We think that in order to achieve trustworthiness in AI, in addition to the requirements already listed in the document, AI-based technology should support the principle of subsidiarity in the handling of all our human affairs, thus promoting decentralisation and weakening currently centralised power structures which severely undermine trust and creativity, which is incompatible with hierarchies and should be the main value that trustworthy AI should promote. The issues described in Section I.5 about the identification, surveillance and mass citizen scoring without consent, about fostering ideological polarisation by means of covert AI systems for the sake of control and manipulation, or the deployment of autonomous weapon systems for conflict resolution could be mitigated by putting an emphasis also on the principle of subsidiarity, which has not been touched in the ethics guidelines explicitly.	This also translates to the requirements and methods discussed in Section II. Data governance, privacy, transparency, accountability, etc. should always follow the principle of subsidiarity, and local communities should always retain the ownership and control of the information generated by them and that is relevant for them to conduct the affairs at their local level.		Information and communication technology should support information flow within communities, but put obstacles to the unauthorised or unconscious delegation of information and control to centralised authorities. AI systems have the potential to significantly support this level of information flow, thus furthering the dynamism, diversity and creative freedom of local communities, which are so important to face the enormous complexity and uncertainty of society. However, the current concentration of data, information, and decision-making in centralised bodies such as the large nation-states and big companies holding almost the monopoly of certain ICT-mediated activities can only become worse by means of AI-based systems, if the principle of subsidiarity is not explicitly supported.

Chiara	Dell'Oro	European Association of Co-operative Banks (EACB)	Before commenting on the requested section, we would like to share some comments regarding the 'Glossary' (page iv).The definition of Artificial Intelligence (AI) in the glossary, which is similar to that used in the AI Definition document, differs from more commonplace definitions of AI. The current definition used in the document focusses on one special part of AI, i.e. 'autonomous systems', which is implemented today e.g. in autonomous vehicles. Of course, 'autonomous systems' are a part of AI, but only one among many others. We think that a better starting point would be a more systematic definition of AI in three steps:1) AI methods including knowledge representation, natural language processing, pattern recognition, machine learning (incl. artificial neutral networks as a subcategory) or machine reasoning. In particular, machine learning ('ML') is a statistical approach to derive (statistical) classifiers based on available data.2) Decision-making systems make use of AI, but – in a simplified approach – consist typically of two parts: a) a classifier (e.g. a credit scoring system with a score value as output); b) a decision rule in the sense 'if then else' to compare a score value against a threshold.The decision-making can be implemented 'manually' – e.g. with a credit (policy) manual to be used by a human credit expert – or 'technically' with a programme and/or software code, which is written by a human and is the technical implementation of the human intention.3) Autonomous systems such as self-driving cars, which react in real time and take actions in the real world, i.e. autonomous systems are decision-making systems (as in 2) with real-time processing. Nonetheless, self-driving cars will follow the traffic code (as pre-defined set of rules). 'Virtual' autonomy is the capability to adapt to changing (real-world) situations in a real-time control loop, but based on (maybe rather complicated) rules predefined by human programmers.It is important to note that none of the above-mentioned systems has an 'own free will' or can make 'individual' decisions, but is always the result of human intention (written as computer code). It is true that human intention is not necessarily correlated to machine output (for example due to errors in the computer code) which highlights the need for robust implementation and testing in order to create trust. However, as we outline in the example below, responsibility remains on the human side.The 'Moral Machine' experiment (see: Edmond Awad et al., Nature, 24.10.2018) highlights the question as to how an autonomous car should 'decide' in the case of an unavoidable accident (e.g. protection of a young bicycle driver vs. protection of older pedestrians). The possibility for a machine to go through a computer programme in real time and take an action in a dilemma does not represent any ethical agency of the car. The question is: who is responsible for these pre-programmed actions?Recently Joanna J. Bryson (see: Joanna J. Bryson, Ethics and Information Technology, Vol. 20, Feb. 16, 2018, pp. 15–26) pointed out this question related to the ethics of human decision-making, but not to the programming of machines [quote]: 'The questions of robot or AI Ethics are difficult to resolve not because of the nature of intelligent technology, but because of the nature of Ethics. As with all	Concerning chapter I, we would like to provide the AI High-Level Expert Group (HLEG) with some brief comments for each of the sections of the chapter.- '1. The EU's Rights-Based Approach to AI Ethics'It would be beyond the scope of this consultation to elaborate on the philosophical relationship between 'rights' and 'ethics'. However, the term 'AI ethics' is misleading. We think that it should be made clear and mentioned in the document that: (i) there is already legislation/regulation, including the General Data Protection Regulation (GDPR), which relates to AI technology; and (ii) the scope of the AI Ethics Guidelines should be 'on top' of existing legislation/regulation, in the sense of specifying the ethical use of AI technology by human beings.- '2. From Fundamental Rights to Principles and Values'It would be beyond the scope of this consultation to elaborate on the philosophical relationship between 'values' and 'ethics'. However, we think that it should be made clear that there is a difference between: (i) the assessment of a technology and the decision of a given society to use it or not (e.g. nuclear power or combustion engines); and (ii) the freedom of will – and freedom of contract – of an individual to make a commercial decision (e.g. buy a product or use a service as offered by a supplier) in compliance with applicable regulations/legislation and based on transparent information made available to the him or her.- '3. Fundamental Rights of Human Beings'As already indicated in the title, this chapter relates to human rights in general (respect for human dignity, freedom of the individual, respect for democracy, justice and the rule of law, equality, non-discrimination and solidarity including the rights of persons belonging to minorities, citizens' rights). Of course, they apply to all human relationships in which AI technology is used by one or all participants. However, there is a fundamental misunderstanding as expressed in 3.3: 'AI systems must also embed a commitment to abide by mandatory laws and regulation, and provide for due process by design, meaning a right to a human-centric appeal, review and/or scrutiny of decisions made by AI systems'. The underlined wording gives the impression that a system based on AI technology can make individual decisions on its own and consequently has to be treated as a new type of social agent. This is a misunderstanding of AI technology in general, as (see above) AI-based systems are always implementations of human intentions and have no autonomy in the sense of an own free will.We suggest taking into consideration our angle by making it clear throughout the text that human beings are the 'ethic' agents in our society and not technical AI systems.- '4. Ethical Principles in the Context of AI and Correlating Values'We think that this this section should be better formulated, as it illustrates the general misunderstanding of these draft Guidelines. Just to give a few examples: • 'AI systems can do so by generating prosperity, value creation and wealth maximization and sustainability. At the same time, beneficent AI systems can contribute to wellbeing by seeking achievement of a fair, inclusive and peaceful society, by helping to increase citizens' mental autonomy, with equal distribution of economic, social and political opportunity.', page 8; and • 'Avoiding harm	Concerning the first point 'Requirements of Trustworthy AI', the focus is changed from a discussion of ethics to a discussion of 'social acceptance' of a (new) technology. Nevertheless, it is a crude mixture of technical requirements (very much standard requirements for any new technology) and misunderstandings about AI technology.As this chapter continues the general misunderstandings about the (human) use of AI technology, we will limit our comments to two examples:- '4. Governance of AI Autonomy (Human Oversight)'The quote 'This also includes the predicament that a user of an AI system, particularly in a work or decision-making environment, is allowed to deviate from a path or decision chosen or recommended by the AI system' makes it clear that the issue is the use of AI in itself, not 'AI autonomy'. Foreseeable implementations of AI technology will be the intention of a natural person ('programmer') and/or legal entity ('company' such as a bank). In the case of a bank, a credit (policy) manual is required by regulations and risk management and has to be followed without any deviation, whether applied by a human credit expert or a technical system. If a rulebook is mandatory, it does not depend on the technology (paper or bits & bytes) that the rule should be fulfilled.However, the responsibility is always with the people in charge, irrespective of whether a credit manual is approved or an AI-based scoring system is commissioned.- '5. Non-Discrimination'The quote 'An incomplete data set may not reflect the target group it is intended to represent' illustrates the general misunderstanding of the Guidelines concerning AI technology.First, this statement holds true for any data set used for a statistical classifier (whether traditional distributions or ANN). That is a well-known problem in statistics, but not specific to AI.Second, there is also the well-known problem that historical data sets will never be 'complete' or 'final' – one always has to find a compromise to optimise false positive/false negative, which has to be an ex-ante compromise and can never achieved 100% for both requirements.The discussion about 'fairness' is rooted in the 2016 ProPublica publication: 'Machine Bias – There's software used across the country to predict future criminals. And it's biased against blacks' analysing the COMPAS software, which forecasts the probability that criminals will reoffend in the US. This triggered an avalanche of opponents and supporters of this 'algorithmic' approach. Nevertheless, Krishna Gummadi from the Max Planck Institute for Software Systems (see: www.european-big-data-value-forum.eu/wp-content/uploads/2017/12/Krishna-Gummadi-Max-Planck-Institute-Discrimination-in-Machine-Decision-Making-EBDVF17.pdf) offered the best summary of the misunderstanding of statistics, as all positions are (partly) right. Both sides used different statistical measures to support their claims of the ethical value of 'fairness'.Concerning the point '2. Technical and Non-Technical Methods to Achieve Trustworthy AI', it applies the well-known 'continuous optimisation' approach to technical systems based on AI technology. Nevertheless, it remains unclear what kind of governance is in scope: the public acceptance of AI technology in general (with	The list provided in chapter III follows from chapter II and, consequently, has the same problems and flaws. We will only highlight three examples of such flaws:- '1. Accountability – Who is accountable if things go wrong?'This is the very usual question of any technical product or service (i.e. liability).- '3. Design for all – Is the system equitable in use?'No technical systems can ever be 'equitable': only the use of a technology by human beings.- '6. Respect for Privacy – If applicable, is the system GDPR compliant?'Of course, any computer system processing personal data has to comply with GDPR.	The draft Ethics Guidelines could be improved by starting with the concept of 'ethical use of AI technology' and a clear understanding of the relationship between AI, traditional statistical classifiers and the use of technology within decision-making processes by natural persons and/or legal entities.Moreover and as a general consideration, we think that the Guidelines should recognise that the many different use cases of AI cannot always be subsumed under the same ethics principles. Such principles should apply predominantly, if not exclusively, to AI systems that are sufficiently complex and/or handle sufficiently sensitive areas. This would be analogous to a risk assessment under GDPR, where a full DPIA is not always required. Obviously, the class with lenient requirements should be much larger than the one with strict requirements. For example, in Android 9, the phone's operating system uses an AI system (based on deep learning) to highlight apps that a user might need next. AI systems recommending which song to listen to next or which ATM should be refilled next, should be exempt from most if not all of AI ethics' principles. Enforcing principles like beneficence, non-maleficence, justice, and explicability do not make sense in the above contexts and could potentially hurt the intellectual property rights of the designers. Only the principle of human autonomy, the possibility to opt-out could be implemented in a meaningful way in the above examples.

normative considerations, AI ethics requires that we decide what “really” matters – our most fundamental priorities.’The introduction section points out the role of trust and ethics concerning AI technology, which differs from a more traditional approach of risk assessment of a (new) technology, as it has been used in academic research and practical use for decades.Therefore, we would like to suggest an alternative approach based on the following three steps:1) Assessment of the technical, social and (maybe) political risks of the use of AI technology;2) Communication of the risks and benefits of AI technology to society; and3) Ethical questions of the use of AI technology by human decision-makers.Such approach would also make it easier to accept the comments raised by some Expert Group members under the section on critical concerns relating to AI. We appreciate why some Expert Group participants might be against including this section – some of the concerns raised are not necessarily directly related to AI as a technology per se, but rather to the way in which the technology is used, to behavioural economics or political choices. This does not mean they do not merit attention from an ethical perspective, but they may not necessarily be translatable into guidelines for AI developers or deployers, which is what the present document aims to do. They would rather fit into the first of the three categories we outlined above ‘1) Assessment of the technical, social and (maybe) political risks of the use of AI technology’.

may also be viewed in terms of harm to the environment and animals, thus the development of environmentally friendly AI may be considered part of the principle of avoiding harm.’ page 9.By reading the two examples, it seems that AI systems are the actors in our current society, whereas all the ‘values’ mentioned refer to human beings as agents in a society, but not to technology. AI systems are – very simply – pieces of technology used by human agents according to their free will.We suggest specifying this concept throughout the whole document to avoid any misunderstanding.The issue of ‘explicability’ is in principle a very technical discussion about the use of statistical classifiers (whether it be traditional scoring based on statistical distributions or classification based on some ‘AI’ pattern recognition). Beside the discussion in research about the explicability of Artificial Neural Networks (ANN) – with very intensive ongoing research work – any decision-making is usually based on an ‘if then else’ programme with a scoring value and a benchmark. Scoring algorithms are always statistical classifiers, which require an understanding of the rules of statistics, including the false positive/false negative problem. It makes little difference for the purposes of the paper whether there is any fundamental difference between a complex rule-based classifier and a simple pattern recognition by ANN.- ‘5. Critical concerns raised by AI’As per our suggestion in our introduction, we appreciate why some Expert Group participants might be against including this section. Some of the concerns raised are not necessarily directly related to AI as a technology per se, but rather to the way in which the technology is used, to behavioural economics or political choices. This does not mean they do not merit attention from an ethical perspective, but they may not necessarily be translatable into guidelines for AI developers or deployers, which is what the present document aims to do. They would rather fit into the first of the three categories we outlined above ‘1) Assessment of the technical, social and (maybe) political risks of the use of AI technology’.From this perspective, there is another issue to consider, in particular under the header of longer term concerns: the scenario whereby only a few very large providers control the market for a given service/product, thereby limiting the ‘free’ choice of customers/users and the possibility to give truly free consent as per the GDPR. Indeed, we should avoid situations where not giving consent could lead to economic/social exclusion. Additional supervisory scrutiny might be warranted.

an approach conducted by public authorities) or the economic enhancement of AI systems in particular (with e.g. ongoing improvement of false positive/false negative ratios as done with any statistical classifier). In other words: what does ‘Trustworthy AI’ mean? The adoption of a technology by society (‘trusting’ this technology) or, rather, understanding statistics (as discussed by Krishna Gummadi)?While the technical methods are rather standard for any information technology (and therefore not required), the non-technical methods do not have the required clarity. We will give one example taken from the sub-section ‘Education and awareness to foster an ethical mind-set’: ‘Trustworthy AI requires informed participation of all stakeholders. This necessitates that education plays an important role, both to ensure that knowledge of the potential impact of AI is wide-spread, and to make people aware that they can participate in shaping the societal development.’ Even ‘normal’ statistics is not understood by the majority of society, even less so when it comes to the false positive/false negative problem e.g. in medical diagnostics (!): the expectation of this requirement is very opaque. It is even more obscure what an ‘ethical mind-set’ should be, as this is not defined and leaves much room for interpretation.

Per Brogaard	Berggreen	Independent	Reference to EU regulationAs the guiding North star is Trustworthy AI and under "Trustworthy AI" - headline "Trust is a prerequisite for people and societies to (ideate, design, my insert) develop, deploy and use Artificial Intelligence". I find it prudent to have some elaboration on the meaning of "Trust" in the specific context. To produce that effect Philosopher K.E. Løgstrup provides us with beautiful and sense creating description (implicitly providing a definition by insight of what trust may be between humans - and perhaps also between humans and AI):" Trust is not of our own making; it is given. Our life is so constituted that it cannot be lived except as one person lays him or herself open to another person and puts him or herself into that person's hands either by showing or claiming trust. By our very attitude to another, we help to shape that person's world. By our attitude to the other person we help to determine the scope and hue of his or her world; we make it large or small, bright or drab, rich or dull, threatening or secure. We help to shape his or her world not by theories and views but by our very attitude towards him or her. Herein lies the unarticulated and one might say anonymous demand that we take care of the life which trust has placed in our hands."(Løgstrup's 1956 book The Ethical Demand, p.18)Scope of guidelineGuiding body creation, Tech specific - alike Ethics Counsel to the Danish Government. It seems ethics is ad hoc and randomly left to or pushed to compliance and legal departments, to do the "Within - out of legal boundaries" check. So, just like we in i.e. ITIL have Change advisory boards, this guideline could propose the establishment of Ethical advisory functions and boards. No doubt for the massive numbers of start-ups and SME government and public function could be put in place to handle this kind of advisory? This should also be seen in the perspective of "a tailored approach is needed given AI's context-specificity", and as such the guidelines are in the nature of cultural formation ethics with strong relations to founding principles of virtue ethics.	Chapter I: Respecting Fundamental Rights, Principles and Values - Ethical Purpose1. "Should" do with technology supports the idea of futurism elements in the guidance and could link to the Ethical advisory function/board. It is also a diversity creation element spanning boundaries of technology and ethics, philosophy and studies of futurism. It is a significant move away for the "just do it" fixer culture, in the sense that it requires imaginary contemplation capabilities of cross-domain competencies.2. The footnote 2) is (to my understanding) central to the guideline use of values in its context, I would consider having it in the text and not "just" as a footnote. In addition, I would suggest that beside referring to the Oviedo convention and the coining of the concept "Ethical purpose", a amount of space be used for clarifying the coining and explaining what "ensuring" EP means (being a key element to Trustworthy AI.3. 3.1 - in several places I miss "mental health/integrity" explicitly. Additional references preferably through direct links would perhaps help the reader to satisfy further interest?4. Direct reference and link to "EGE" and AI4People (Springer provides open access). We touch here a cardinal challenge - open access to information and knowledge vs. IP/research & knowledge ownership and whether institutions like EU could/can/should make a buy and release of given reference research e.g. Ibo van de Poel referenced material. To some extent this relates to the "informed consent" of the common public (like myself) and secondly it has limiting effect on knowledge sharing and thirdly enhances the autopoietic effect in academia and fourthly a general negative effect on the perceived relationship between academia, business and public world - in a sense it showcases the opposite of inclusiveness. Yes, I do know it is largely a question of the free enterprise, the right to make money, commercialism… vs. The right of information and knowledge of the common public - after all a lot of research are done and produced in part publicly funded universities and research institutions!, by European standard. The question is of course (oversimplified) - how will I as a commoner ever get smarter, better informed and hence able to provide informed consent if I'm "excluded" from accessing information, knowledge, wisdom freely. This is no insignificant problem in the free world.The whole idea Ibo van de Poel provides in regards to technology as social experimentation could actually provide cardinal to how we approach AI experimentation and even perhaps public testing. Further, it could (hopefully) bring more attention to abduction as a method and hence reduce the conservative attitude of clinging on to deduction and induction as the only true academic methods. In light of recent decades rise in complexity research, there may be still some hope.5. If anything, this section provides amble argumentation for the need of cross-domain/field ethical considerations and guidance along with the Oviedo convention. For sure AI is in itself important and potent and in the cross-use and application this will be even more so.	Chapter II: Realising Trustworthy AIThe mapping methodology (Ibo van de Poel et.al.) seems central to this chap. And section 1 - so much so, that description perhaps in an appendix/addendum could be relevant. I personally use and reuse a mental model design over the years based on experience (many would argue that "virtues" does not belong in the model, I do NOT agree - as the virtues are the guidance of the inner being in the conversation to/with itself). (Not able to put in the model graphic!!!!!) Would it be worth the effort to try and make the mapping visually and through the use of axiology's (considering the loss of subtle details but compensating through text) - alike mind mapping, relation diagram et.al.? The purpose is the creation of overview (even simple tables relating to chap. 3, a readers traceability and relation overview (ease of use).One thing I miss is explicitly mentioning of "Mental" safety and health along the general "Safety" and physical health/integrity. It speaks to trust and the "true" as a concept, and hence of utmost importance to the guiding North Star of "Trustworthy AI" and Human-centric approach.Concepts of ethical hacking seems a relevant point to mention, along with a prospect revitalization of "attribute" methodology and technology to ensure transparency, traceability, explainability and explicability. In general, I think the success is highly dependent on education/formation ("bildung" in German, "Dannelse" in Danish) - and that has significant impacts on and outside the field of AI. A main concern is the lack of interest and a general position of indifference - well put in the words of Aristotle: NICOMACHEAN ETHICS"These two rational faculties may be designated the scientific faculty and the calculative respectively; since the calculation is the same as deliberation, and deliberation is never exercised about things that are invariable so that the calculative faculty is a separate part of the rational half of the soul."Which also calls attention to inclusiveness, diversity, informed consent, readily accessible information and knowledge - if we feel we are at loss against the "System" be it government or capitalism - we lose interest by the rationale, that we do not stand a chance against the Giants of the world then we stop deliberating, because we see things as being invariable. Hence also the relevance of technology as social experimentation! (K.E. Løgstrup, The ethical demand, p.10 (my translation): "Faith/belief without understanding is not faith/ belief, it is coercion".Let's face it, Law can regulate the world, it can even support the government of the world, but it cannot rule the world - only ethics and moral can do that, and that is a human endeavour and because it is a human-dominated world. When contemplating the unknown, we cannot regulate - we can sense and probe firstly and try to imagine scenarios. Let's not govern by rule before we have seen what emerges, and let us be present in the emergence by way of e.g. inclusiveness, diversity, openness, sharing, debating, discussion and dialogue based on the human rights and our inherited culture and ethics.Standardization is much appreciated in the field of practice, also in applying ethical consideration. On the other hand, it has a tendency to make things implicit, taken for granted, less dynamic & closed to	Chapter III: Assessing Trustworthy AIScenarios of potential system and data manipulation and mitigating/preventive built-in designs and human controlling and auditing.How are manipulation monitored on a continuous basis in the aligned cycle with security issue checks, system and data updates and regular system execution checks?Has processes and responsibilities been put in place to follow field research for AI technology and societal impacts?Has potential influence on user behaviour been assessed and described (e.g. in regards to recommendation systems, nudging….)?Have SOPs been put in place in case of suspicion of malfunction (bugs), system and data manipulation, security breaches etc?	Terms used needs to be coherent, as IT/Tech person the lifecycle of technology is more or less consistent: Innovation = Ideation/Invention + Design & architecture -> Development -> Test -> release/deploy etc. The Ideation/Design is left out sporadically and too often. An example lifecycle in appendix could be an idea. Executive summary:In principle the claim that AI benefits outweigh the risk is unsubstantiated at the present - it can maybe be predicted. Secondly, it is hardly an argument - the situation is that the development/progress cannot be stopped, and hence also from that perspective, we need to maximize benefits and minimize risks. Love the structure in 3 layers of abstraction - brilliant and easy to understand and makes the draft coherent - solid framework, thanks.I urge again the use of visuals (models, diagrams etc.). Especially on the attempt to create further understanding within the field of Ethics. We must remember that we to a certain extent are "pushing" and attempting to merge two very different fields (Ethics and technology). Explaining interrelations, the interconnectedness of a) the abstraction layers and 2) fundamental rights and fundamental principles and so on, in models have an appealing effect on natural science brains (I know as I come from both worlds). In addition, we should not underestimate the future communication of content, and also be aware and focus on how messages can be relayed to different management levels and decision makers (from middle mgmt. to C-level). Coherent models supported by competent presenters will be required - most likely by some sort of cascading principle with anchor points in different communities incl. HLEG AI and EUROPEAN AI ALLIANCE etc.

					the evolvement and development on both the technical and ethical sides of the assessments. The management of standards must hence be kept dynamic, and also include a meta-level preferably including resources from all aspects of field e.g. government, business, research, users….		
Herve	LE GUYADER	ENSC (Ecole Nationale Supérieure de Cognitique)	The framework articulated by this document for Trustworthy AI relies on two critical components: 1) ethical purpose and 2) technical robustness. I would like to propose a third component, "social acceptance", which, IMHO, is not merely a consequence of a successful combination of the two former components, but rather a goal in itself. Social acceptance implies explainability, equal accessibility. In very broad terms, such a three-legged strategy is a "classic" when it comes to looking for optimising technology's potential positive impact on society as a whole: one needs a combined approach where (1) Private sector, (2) Public sector, (3) Civil society work together and complement each other's efforts. Regarding the "Human-centric" nature of the approach: Human-Centric AI should also explicitly set the balance between Autonomy, Delegation and Responsibility. This balance is to be defined according to the different categories of possible use cases and is to be based, in particular, on corresponding risks.Regarding the necessity for Trustworthy AI: AI also has a strong bearing on security and defense matters, whether it's for protecting personal data, fighting terrorism, protecting our societies' assets from cyber-attacks… It also plays a key role in influencing our citizens, in their shopping habits, in their access to and personal expression on, democracy, in setting and updating their personal set of values and beliefs.Regarding this particular draft document, "AI Ethics Guidelines": Once it's finalized, it could make sense to encourage the creation of some more operational application, domain specific renditions of it, perhaps starting with a handful of EC departments and executive agencies.	Regarding the "The EU's Rights' Based Approach to AI Ethics paragraph", I would first offer two main comments/suggestions: 1. A proposed list of five key principles:1.1 Respect of fundemental rights (ab initio and in itenere),1.2 Non-discrimination,1.3 Quality and security (data, models, algorithms, interdisciplinary approach, security (including cyber)),1.4 Transparency, neutrality, intellectual integrity (accessible, comprehensible, auditable),1.5 User control (including UX, but also need to be informed that AI is being used, opt out option?).2. IMHO, this document should at least touch upon the following issue:Europe is a fairly unique democratic diverse, albeit united, platform that can probably enact AI ethical guidelines that will have some binding consequences for (mostly European), public and private parties.Other regions of the world may act differently and progress on AI matters with different views, guidelines and agendas. This is true of governments, this is also true of private companies. Both, thanks to the power and global appeal of "digital", can easily access, affect EU citizens, EU private and public bodies. In other words, the document should acknowledge that this is a global, open world we live in, especially prone to vulnerability caused by "accepted digital contagion", that of course includes tools and services that now benefit from some "AI magic".This (the point made above) is the potentially negative aspect of living in a globalized, open world, but there's a positive aspect to it:GDPR is a good example of a European-led initiative, with bearings on any international company that would use European citizens' personal data that MAY represent a competitive edge to EU developers. It's worth reflecting upon the competitive advantage an ethical approach to AI COULD bring to European AI developers.Regarding the "Ethical Principles in the Context of AI and Correlating Values" part of the document, more precisely its two first principles ("do good", "do no harm"), and very sorry about this (it's always so much easier to comment than to create …), but I fear there's some naivety factor to be taken off the two corresponding paragraphs, just because every man-made system, a	FWIW, working on this contribution to the HLEG initiative often reminded me of the study I did for the EC a few years ago on "eInclusion public policies in Europe". I'm not sure how easy it is to find this study on the many EC web sites, but I would be happy to send a copy of it via some electronic mean and/or to share experience with the Group.Some strong similarities emerged between your "Realising Trustworthy AI" and the "Realising eInclusion" element of my study, such as:(i) The absolute necessity to have civil society, together with public sector and private sector, involved into this "realization",(ii) Accessibility and usability (identify and address the possible "divides" within society when it comes to, in this case here, AI,(iii) Need for leveraging any possible binding rule to encourage and enforce, here, ethical AI,(iv) ...Maybe there is some inspiration to be found in this study. Again, for what it's worth, here is the final set of 6 public policies for eInclusion that I suggested taking a look at and adapting to each Member State's idiosyncrasies, from an initial set of 12. First one (Appointing a coordinating authority) may sound a bit bureaucratic, but keep in mind the UK had their full-fledge eInclusion Minister for some time back then…The second one (Awareness raising …) could have some mileage in the AI context, as it includes the use of "Champions" (championing the cause, in this case, of ethical AI) … Policy # 1: Appointing a coordinating authorityPolicy # 2: Awareness raising, Stimulating and supporting initiativesPolicy # 3: Designing a specific eInclusion strategy vs. Mainstreaming eInclusion into traditional policiesPolicy # 4: Enforcing eInclusion public policiesPolicy # 5: Addressing specific excluded groupsPolicy # 6: eIncluding the territory per se within a globalised worldThen, regarding the "Ethics & Rule of law by design (X-by-design)" paragraph, the whole "by design" approach is strongly reminiscent of GRPD rules (with its "privacy by design" rule) and it's quite tempting to work on that analogy, but GRPD's potential effectiveness/success relies on 2 very strong "sticks":a) there is a regulatory body in each Member State,b) sanctions shall be applied	This chapter identifies 10 requirements for Trustworthy AI. Although it can be seen as related to several of these existing requirements, I find the "Bias" issue to be a key one, calling for some necessary "Bias detection/denunciation" additional requirement. Just like, because of the growing percentage of "fake news", most "serious" media now have a "fact-checking" section, I would encourage a strong impetus for "Bias detection/denunciation".Bias can indeed exist at the data level (training data, for instance) but it can also, intentionally or unintentionally, be introduced at the reasoning phase. Intentionally, if there is a will to exploit the "appeal" of AI to obfuscate bias in achieving a goal (e.g. discrimination). Unintentionally, for example, if the designer isn't capable of an impartial, rebuttable reasoning because of his/her beliefs.Bias can also be introduced at the algorithm level, again intentionally or not. Also, assessing accessibility would be, IMHO, important to add to this list of requirements. For the very reason AI can bring substantive benefits to individuals and society, equal accessibility to AI is essential. "Fair trial" is one of the fundamental principles of Justice. If one party can access (i.e. afford) AI while the other cannot (for preparing trial, weighing options, finding jurisprudence, the trial can't be fair.	I would first like to commend the HLEG for drafting this document and wish them the best for exploiting the feedback they will hopefully harvest from this consultation. Many suggestions I was thinking of putting forward when starting reading this document became unnecessary when getting to the following page/s, and more were deleted after a second reading!Should this be deemed useful, I'd be more than happy to share more thoughts with the HLEG. As I wrote above, some of the work I did for EC on eInclusion might be relevant to this ethical AI study.Regards,Hervé Le Guyaderhlg@ensc.fr

				fortiori an AI fueled one, is capable of both "do good" and "do harm" (Autonomous car would spring to mind).Regarding paragraph “5.5 Potential longer-term concerns”: One way of categorizing AI is to identify three broad categories within it: AI can be (1) Augmentative (helps a human achieving more, in a Man-Machine Teaming or “HAT” situation), (2) Substitutive (where AI “replaces” a human in a particular task), but it can also (and, probably, will become more and more) (3) Hybrid, i.e. AI becoming physically embedded within the human (rather a Cyborg by then), questioning the very definition of what a human being is. My impression of the current AI ethics document is that it has not taken stock of this third category. IMHO, this is a strong longer-term concern.Slightly less futuristic, the document also may want to look at this angle: The typical use case dealt with, or implied here, is: One person using/being exposed to One AI based system. In reality, the person, in most cases, is not alone and interacts with other persons in accomplishing his/her task. By the same token, the AI based system is not functioning alone, by itself, but is more likely to be immersed in a (AI) system of (AI) systems. Just think Internet of Things, most of these "Things" have some form of embedded AI with distributed delegation and autonomy. The issue tackled in this document (ethical AI) should acknowledge this somehow.	in case of non-compliance.None of this exists today re: ethical AI.Regarding the necessary « Diversity » called for in the document, I suppose it (also) means "interdisciplinarity", identified as a pillar, key factor for AI. For reference, the French Official Journal of 18th December 2018’s definition of AI reads: “A theoretical and practical interdisciplinary field whose purpose is to understand the mechanisms of cognition and reflection, and their imitation by a physical and software device, for purposes of assistance or substitution to human activities.”Open data: IMHO, the document should mention the need for stronger Open data, especially for public sector (generated) data and for the necessity of its use for machine learning devices.Lastly, trustworthy AI indeed relies on the two first components the document suggests, i.e. (1) its development, deployment and use should comply with fundamental rights and applicable regulation as well as respecting core principles and values, ensuring “ethical purpose” and 2) it should be technically robust and reliable, but also relies on a third, equally important pillar: explainability.		
Carolina	Brånby	Confederation of Swedish Enterprise, representing 50 sector organisation members and 60 000 member companies in Sweden.	A. The uptake of AI technology is highly relevant to business competitiveness and capability to innovate improved goods and services. Not least for addressing challenges in society like climate change, productivity and healthcare. We welcome the objective of the EU HLEG to develop Ethics Guidelines which are not just a compilation of values and principles to be respected, but which most importantly aim at providing guidance on how to actually implement these. The Guidelines for trustworthy AI could strengthen the uptake of AI, but then it must be relevant, meaningful and concrete. A, Para 5-7 (and Executive summary): “…Trustworthy AI…requires ethical purpose…to enable responsible competitiveness, as it will generate user trust and, hence, facilitate AI’s uptake.”, “…position itself as a home and leader to cutting-edge, secure and ethical technology.”, “…we will fully reap the benefits of AI.”Firstly, we agree with the statement that trust is a prerequisite for people and societies to develop, deploy and use Artificial Intelligence.However, calling ethical purpose introduces a bias from the outset as it seeks to direct innovations. The guidance should aim at providing a framework based on ethical principles where design is ethically aligned but not limited by purpose.Also, the presumed correlation between competitiveness and an ethical approach to AI might very well be true, at least in the long run. However, one could also envisage that an approach that might not be as structured and “ethical “as the one outlined in the paper will create competitiveness in the shorter perspective and due to the often-inherent characteristics in the digital area of benefits of scale, first mover advantage etc. This might also lead to competitiveness in a longer perspective even	Calling Ethical Purpose introduces a bias from the outset as it seeks to direct innovations. The guidance should aim at providing a framework based on ethical principles where design is ethically aligned but not limited by purpose.I.3 Fundamental Rights of Human Beings I.3.4 Equality, non-discrimination and solidarity: A clarification might be useful – what does “a fair distribution of the value added being generated by technologies” actually mean? Is it fair access to the use and benefits of the technology that is intended (e.g. possible access to medical innovations based on AI), or is it a more “fair” distribution of funds/profits generated by the technology? Our presumption is that fair access is intended. More a question of wording in the next sentence, workers and consumers are hardly minorities but might be disadvantaged compared to other groups in society in other terms (economically, access to information etc). I.4. Ethical Principles in the Context of AIPlease explain better and shorter the aims with I.4. Ethical Principles. Morality and ethical question cannot be answered, therefore the five “principles” should be taken out. How should these five be handled? Is there a hierarchy in between them? Of all fundamental rights, are those the ones that should be emphasized when deploying AI? The most guiding bullet in practice is Explicability: Operate transparently (the mixture of fundamental rights and principles are confusing). I.4. Bullet 2, “Do no Harm”: The paper might benefit from a bit more precision regarding the limits of this principle. International and domestic law is pretty clear that in certain situations (e.g. national security, to protect life and health, environment etc.) there might be legitimate needs to overrule this principle. There certainly are cases where	The basic idea of the chapter is good but could be more concrete. The design of the chapter in its present form is more problematising than substantialised. II.1. Requirement of Trustworthy AIThe ambition is good describing which areas to think through when working with AI. It would be useful with less extensive and not too general text. Maybe some points can be combined: 2 + 7, 3 + 5, 8 + 9.We suggest testing the 10 requirements against reality - small companies as big companies- to see if they can be compliant. II.1.3. Design for all: “Systems should be designed in a way that allows all citizens to use the products or services regardless of their age, disability status or social status.” “Design for all implies accessibility and usability of technologies by anyone at any place and at any time ensuring their inclusion in any living context. Taken literally, this is an extremely demanding challenge which might be hard to live up to. There has recently been a similar discussion in the EU regarding a proposal for a “European Accessibility Act” for products and services which is still not concluded. From a business perspective there were several reasons for hesitation regarding this proposal, one being the possibility to live up to very lofty political goals that anyone regardless of physical or mental limitations should be able to use any product or service (not to mention the costs this would imply).II.2. Technical and Non-Technical Methods…The areas described are already applicable and in themselves nothing new except maybe the part describing traceability. II.2.1: "To tackle the challenges of transparency and explainability, AI systems should document both the decisions and the whole process that yielded the decisions, to make decisions.""For a system to be trustworthy, it is necessary to	The idea of having a concrete assessment list is very good and helpful.We call for a clarification that underlines that all 10 points in the assessment list cannot and does not have to be operated by all companies because of different requisites, see our comments on the 10 requirements of Trustworthy AI above.III.1.AccountabilityThere is sometimes more than one actor responsible when things go wrong with AI. This need to emphasize in the chapter of accountability. We call for a deeper account of the lifecycle. Many times, there are AI users that are neither the developer nor the ultimate user. Not all companies have the resources to establish a review board of ethical AI. We understand that the document in general has a broader perspective – but we call for a clarification that underlines that all these points in the assessment list cannot and does not have to be operated by all companies because of different requisites.III.3. Design for allDoes the system accommodate a wide range of individual preference and abilities? What does a wide range mean? It differs a lot depending on data governance and the size of the company. If the system must be special designed for all special needs or disabilities it could be very costly to the provider, and sometimes even impossible. III.5. Non- discriminationThe non-discrimination point should not be phrased so they hinder a company from contractual and industrial freedom. For example, costumer segmentation is something many companies do today with AI, will this not be compliant with the guidelines? GDPR is regulating how companies and authorities may use personal data.III.7 and 8. Respect for Privacy and Human AutonomyIn the matter of information, it can not only be the responsibility of the companies that are	Swedish Enterprise welcomes this initiative and we do support the important aim to raise AI ethical awareness. The Guidelines are needed to create better understanding about upcoming issues concerning further uptake of AI. From business point of view the Guidelines have a significant signal value to contribute to consumers information and trust. The Guidelines are by now quite general, academic, educational and high level but when the content gets more concrete it could be practically applicable on how to act ethically. Now highlighting ethical aspects in a specific area within the digital does not facilitate business or innovation but provide a guidance on how to actually implement these.There are some concerns. Swedish Enterprise call for an Impact Assessment about how the approach with ethical purpose will affect innovation in Europe, competitiveness and business uptake of AI.Also, to sign up - this creates one of our biggest concerns. What does it entail to endorse and sign up to the Guidelines? More information needs to be taken on- board to explain the appliance to be compliant. Moreover, it´s important to be aware of and discuss if the Guidelines in practice will become volunteer or mandatory, e.g. in public procurement requirements? The large compliance costs are always a risk for businesses (compare huge costs for GDPR). Business needs a much more actionable guidance! Shorter and more focused document. For our lager companies it’s crucial to have Guidelines at international level. In this regard, please note the short and handy OECD guidelines. We believe it could be a good idea to split the Guidelines into two different documents. One educational part for all stakeholders and one part consisting of Realising and Assessing Trustworthy AI. Further regulation on AI and

though a structured and ethical approach in that case will probably come later and more of an afterthought than as a precondition. Point being that if Europe is going to catch up in the international race for competitiveness in this area, it is important that the business conditions are right and that developments are not hindered by unnecessary obstacles.Swedish Enterprise call for an Impact Assessment about how this approach will influence Europe's competitiveness and business uptake of AI.Purpose and Target Audience of the Guidelines, para 2: Sign up -what does it entail to endorse and sign up to the guidelines? More information needs to be taken on- board to explain the appliance to be compliant.B. A framework for trusted AIIn conclusion, section II and III as such are far too extensive, and the recommendations will be far too time-consuming for companies to apply to their day-by-day operations. It is also the case that the document in these sections, by making theoretical assumptions and pointing out purely operational measures, makes it difficult to distinguish what is important and less important. We call for a risk-based approach where the Guidelines serve as a tool to enable the users of the Guidelines to ascertain the risks derived from their particular scope, circumstances, technology and impact. The mitigation must be proportionate to the potential adverse impact. Figure1: We are convinced that the figure in the document describing the framework for Ethics and AI in itself can be of use for professionals developing various AI solutions by showing a whole to consider when producing AI services.

individual rights and liberties might be legitimately compromised in order to protect the rights, interests and liberties of others. Since these guidelines are intended to be an instrument that different stakeholders can endorse, it is important that as much clarity as possible is achieved to avoid disputes over what constitutes reasonable interpretations of the text. I.4. Bullet 3 The principal of Autonomy: "Preserve Human agency": We question "users" right to opt out and a right of withdrawal from AI systems at working places. We strongly oppose to footnote number 13. It would disproportionately restrict the Freedom to conduct a business in accordance with Community law and national laws (EU Charter of Fundamental Rights, article 16) and hamper European competitiveness.I.5. Critical concerns raised by AIThis part consists of examples of pernicious use. Maybe it could be moved to an appendix or deleted?I.5.1 People suffer from "consent tiredness" and agree to everything to be able to purchase goods and use services. Good that all legal grounds in GDPR are mentioned, not only consent. Contract with the data subject and legitimate interests are more secure to use than consent in article 6, GDPR.I.5.2 Covert AI – We suggest this part should be exchanged for Embedded AI and focused on influencing and nudging.I.5.3 It's both very complex and costly to ensure opt-out options in AI-systems. The regulation that secure dataprotection is GDPR. You must have at least one legal ground to use personal data. If the only legal ground is consent you always have a right to have your data deleted.I.5.4 AI applied in weapons systems. What is described is an autonomous system without human control over the critical functions. However, there might be less draconic and far-reaching applications of AI in weapons systems that will certainly have the potential to bring harm to individuals. However, the UN Charter (art 51) gives countries a right of self-defense against armed attacks and hence, with reference to the above, it might be a good idea to insert some text that makes clear that there are exceptions to the principle of "Do no harm". I.5.5 Footnote 18 - recapture should be added as an appendix

understand why it had a given behavior and why it has provided a given interpretation." We do believe traceability would be very good to achieve trustworthiness, but perhaps not quite easy to apply. II.2.2. Non-technical methodsCodes of Conduct: Sign up - this creates one of our concerns. What does it entail to endorse and sign up to the guidelines? More information needs to be taken on- board to explain the appliance to be compliant.

using the AI-systems to describe technically parts and how the processes are taking place. Companies must be able to rely on a certain amount of prior knowledge of the consumer that has developed during education and thru the government information. Otherwise the information burden will be too heavy on companies. There is always a risk of using wording like "clearly communicated" - what does that mean?

ethics may create unintended problems and limit the business ability. There are new technologies emerging and all frameworks should be technological neutral as far as possible to not hamper competitiveness and add regulatory burden on companies.It would be helpful to acknowledge the framework already put in place to make technology used in an ethical way, like data protection rules, liability rules and rules within the healthcare sector for example. And therefor, in order to avoid redundancy and the risk of contradiction that this engenders, the document should not state things stated elsewhere (in other pre-existing bodies of law, regulations, etc.). Instead, references to these sources should be made.

Anonymous	Anonymous	Anonymous	IntroductionCommerzbank welcomes and supports the acknowledgment in the introductory sections of the Guidelines that this is an emerging area and that changes to the Guidelines will be necessary over time. The start of the guidelines also recognizes that different contexts will require different approaches, with flexibility required in application (page 3, "Scope of the Guidelines"). While acknowledging that different contexts require different approaches it seems crucial to not only ensure technology neutrality of ethical standards but to also make sure that a level playing field across different industries exists with regard to the ability to use AI.1. There is no clear specification of function of the ethical guidelines. It should not be an administrative set of guidelines, but a motivation to create AI and communicate on ethical values related to AI. The third pillar of the Commission's vision "ensuring an appropriate ethical and legal framework to strengthen European values" moves the objective of an ethical guideline and a legal framework very closely together, which is not appropriate in this case. We should separate the ethical framework from the legal framework.2. The missing specific definition of a target audience (either public administration or private business) is a big source of dilution in this case. For public institutions and the administration it is important to be transparent. As a private business, on the other hand, some processes are likely to be hidden due to concerns of privacy and competitive advantage. This provides the justification of other rules, such as consumer protection, product liability or antitrust law, which have to be specified and further developed to cope with AI solutions.3. No clear definition of the object of the guideline: Are they specific to AI - then conventional concepts such as technical failure, user liability, etc. are completely sufficient – or is it an unknown concept of digitalization in the far future – then the guideline is not relevant at all. The definition of the object sets the focus on technical descriptions.4. The rationale, implications and foresight of the guidelines, which are initially of a voluntary nature, cannot currently be assessed. According to our understanding, the high requirements for transparency, explainability, comprehensibility and non-discrimination of AI systems in particular could have a negative impact on broad economic use. What are topics of regulation and what are topics of ethics-communication?5. There is no adequate adaption of data-issues and Deep Learning for shaping explainability and trustworthiness. It is difficult to identify and qualify when a model is homogeneous or when it is biased. There is still some work to be done on data quality - a clear metric that shows how the data set affects the model.6. No clear adjustment to a concrete ethics approach: Is it a modern ethics of communication and virtue or an obsolete natural rights issue? Is it utilitarian? Do they focus on a common good and a "summum bonum" of the society?7. Ethics Guidelines are not a solution but the beginning of a process and discussion and should be separated from all legal or soft-law ambitions. It is important to note that ethics should be technologically neutral at all times.	The Principle of Non maleficence: "Do no Harm""Harm" needs to be defined in more detail. Does harm apply to all industries alike? Is harm done by a self-driving car different or the same as harm inflicted when a loan is not granted? Regarding the banking industry, there is no harm added or deleted when using AI models for example to increase efficiency in operations. Deciding whether to grant a loan to a customer is the standard business of banks ever since. A customer who is not able to pay back a loan might even be better off when not being granted a loan, thus avoiding excessive over-indebtedness. Banks have always used some sort of scoring to evaluate whether a customer is able to pay back his or her loan. Traditional scoring models are oriented to past salary pay back habits etc. New developments, for example in China, show that there is a trend to social scoring where it also matters who you are friends with. Preferably, this will only be allowed under cautious constraints and boundaries as in these cases full transparency can never be granted without revealing personal data of other customers. This is undesirable. As already outlined in the GDPR the customer shall have the right of access to get meaningful information about the algorithms' logic involved and the right to obtain human intervention.The Principle of Justice: "Be fair"It is impossible to ensure that all individuals remain free from all kind of bias. Bias is inherent in all societies and systems, regardless of AI. As a consequence, historic data sets used for training of AI systems also include biases by nature. This is not only true if the training data comprises of historic human decisions. If it is obtained from historic events (e.g. credit defaults) then AI goes beyond human observation biases. This is exactly what has been observed in the Amazon recruiting engine. It was a hiring tool that aimed to identify the best candidate for the job, but turned out to prefer men over women as in the historic training data good software developers where actually mainly men. It should also be noted that it might sometimes be difficult if not impossible to detect unintended biases or discrimination. Detecting unintended biases would, for example, require to collect sensitive attributes protected by law (such as age, gender, race, religion etc.) in order to ensure that no correlations to other attributes incorporated in the model exist. This would not be socially acceptable. The main problem here is that as soon as the discriminating attribute is correlated with the target variable all other attributes that are useful for predicting the target variable will also be correlated with the discriminating attribute. Hence, it is statistically impossible to avoid discrimination a priori. However, averaging over discriminating variables a posteriori completely removes unintended biases. However this requires the discriminating attributes to be defined and cannot be done if the discriminating attributes are unknown (e.g. sexual orientation). Coming back to the example of the Amazon hiring engine, it has been reported that the engineers deleted the attribute of gender and retrained the model. Still the algorithm preferred men, as correlation was present in the remaining data, e.g. women's colleges or all-girls' schools were handled as a refusing criteria while preferred candidates used verbs like	It is important to understand that AI models indeed simply model complex reality. They cannot be expected to "explain" complex reality in a scientific sense. For example, economists have tried to explain financial markets or predict GDP, inflation and other economic aggregates for many years without ever succeeding completely due to the highly complex nature of underlying data. AI models strongly rely on the observation of inter-correlation between attributes which are too complex to be easily understood. That is, model validation is the only way to establish trust in AI-based tools.We are convinced that sustainable customer relationships are built on mutual trust. This also involves indicating when a customer is interacting with AI, i.e. a chatbot or interactive voice response or a human being. This transparency with respect to different communication channels is a crucial element of mutual trust.		AI models give rise to new options for companies to analyze their customers' demands and create tailor-made solutions. This can be done with or without consent of the customer. However, exactly this question of what can be processed with or without customer consent is detailed out in the GDPR. Especially when it comes to fraud detection the AI toolset is of utmost importance, as the volume of data and interconnectedness is extremely large. Moreover, the screening must be done in such short time frames that efficient and effective automation is the only option.

				"executed" and "captured" which happened to be used mostly by male candidates. A solution to the problem would have been to keep the gender variable and run the decision algorithm twice once for "female" and "male" candidate and finally average the result.The Principle of Explicability: "Operate transparently": We agree that transparency is key to building and maintaining citizens' trust in AI systems. However, the document presents transparency as always desirable. We believe transparency should not be so detailed as to undermine the use of AI in certain circumstances. It is indeed crucial to find the right degree of transparency vis-à-vis individuals, competent authorities, jurisprudence, etc.! Transparency goes hand in hand with a loss of intellectual property and must therefore be well balanced.		
Miguel	Oliveira	Portuguese Psychologists Association	The different countries that develop AI come from different historical, social and cultural perspectives, as well as from a perception of human rights that is not universal. It is therefore necessary that AI systems do not reproduce the same biases and prejudices that result from the historical, social and cultural differences of their countries of origin. Robot Sophia is an example of this, since she is a Saudi citizen and has appear to have more rights than women in Saudi Arabia.	In section 3.1. Respect for human dignity we consider it would be adequate to add the idea of helping people to make their own decisions. In section 3.4. Equality, non-discrimination and solidarity including the rights of persons belonging to minorities we thought it would be useful to discuss equity and non-equality. In section 4. Ethical Principles in the Context of AI and Correlation Values, we would like to comment that the theory of Beauchamp and Childress is based in the same paradigm of the Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine. In an absolute conflict between the individual and the society, the individual should be first considered. We think that this is true if we are talking about the fundamental rights. In that way, and talking about these principles, we think it should be added, the principle of responsibility - taking in account Hans Jonas perspective, and including a social responsibility principle, and the principle of integrity, a very important one to state that everyone should considered all these principles before undertaking any kind of options. Still in section 4, and about The Principle of Explicability: "Operate transparently", we consider that responsibility requires helping others to reach their own nature. AI should be responsible for promoting every person's autonomy and well-being. This autonomy is based on increased self-knowledge, which will help and empower people to make more (own) conscientious decisions. Responsibility is also linked with the conflict between personal and society interests. The interests of society, as well as the interests and rights of each individual, must be taken into account. The difficulty with that is that often-individual interests collide with societal ones. In this case, AI must try to eliminate the potential negative consequences for each and try to find the best possible outcome. Nevertheless, it should be clear that the	The ten requirements for a responsible IA should include education, the promotion of digital literacy through a strong commitment to the development of digital skills that enable all citizens to understand AI. Citizens cannot constitute a passive subject in the dialectic with AI and be only someone who believe and trust. It is necessary to empower the citizen, with a set of skills and tools, so that they can exercise their active citizenship also in these matters. In section 4. Governance of AI Autonomy (Human oversight), we consider it should be clear who is the responsible about the results. And we should try to avoid the growing idea of getting the right or wrong answer just after the data reading. This can develop a stricter feeling of right and wrong and diminish the idea of flexibility so important in our social organization. In section 6. Respect for (& Enhancement of) Human Autonomy, we think that careful is needed not to sell the ideal that these machine approaches always represent the best option. In section 7. Respect for Privacy, we reinforce the need for institutions (like Banks or Unions) to keep the data safe.	The chapter starts by stating that "The primary target audience of this chapter are those individuals or teams responsible for any aspect of the design, development and deployment of any AI-based system that interfaces directly or indirectly with humans, i.e. that will have an impact on decision-making processes of individuals or groups of individuals". In this sense, it is surprising that there is no reference to Psychological Science and all its accumulated knowledge about the processes of interaction between AI and citizens. We think it should be added to the Assessment questions the following: • In section 2. Data Governance: Who is keeping the data? • In section 3. Design for all: And between countries? Is this technology to be shared by everyone? • In section 6. Respect for Privacy: It is possible to grant total privacy? How to explain this to people? • In section 7. Respect for (& Enhancement of) Human Autonomy: we think that more important than the way we explain it is important to try to evaluate how people understand.	In general terms, it is important to point out the scarce participation of specialists in Psychology, taking into account all the concepts and psychological processes involved and the contributions that Psychological Science can bring to the AI area (in fact there is no psychologist in the Group that makes up the High-Level Experts Group). It is also worth mentioning the diminished emphasis on education and the promotion of digital literacy in AI of citizens, in particular as opposed to the concept of Trust (trust/trustworthy is referenced 125 times throughout the document, while the words education or skills appear only 7 and 5 times, respectively). Specifically, we consider that it is imperative to address the Human Enhancement issue, that it is not addressed in this document. Similarly, it is also necessary to address the issue of decision-making – how AI can interfere in the way people usually make decisions. We think the differences in the world and how AI can affect the relations between countries and powers should be another topic addressed, as well as Big Data as one of the present big challenges for world governance. Finally, we consider it useful to extend the process of analysis and decision making of documents such as these, implying a greater diversity of participants in the process of defining essential concepts for the later streamlining of processes.

				individual should come first. Integrity implies coherently applying these ethical principles of AI in order to make it more and more accessible to the general public. As such, integrity helps to promote acknowledgement and trust in the profession. Therefore, integrity as defined might be compromised whenever some agent allows to be influenced by his/her/it own interests or beliefs. In the end, one must pay attention to potential conflicts of interest, which at a later time may put the AI in the position of having to disrespect the ethical principles, even if involuntarily. About section 5.3. Normative & Mass Citizen Scoring without consent in deviation of Fundamental Rights, we believe that this is a complex issue. We already use scoring with, for example, employees. The public often score a person after shopping, and it is difficult for the employees to say no. In the end, bad judgements have bigger impact then good judgments and judgments are influenced by diverse factors. So, we would agree to fully don't use people scoring. In section 5. Critical concerns raised by AI is stated that "AI systems should be developed and implemented in a way that protects societies from ideological polarization and algorithmic determinism." There is the problem of receiving inputs according to what we "like" and the people we "like", and therefore the fundamentals for the decision-making process are fallaciously well grounded, since it comes from a reduced spectrum and with little contradiction and, therefore, may cause biases. It is also important to generate news forms for giving consent. In section 5.5. Potential longer-term concerns, we are concerned about several topics: 1) Big Data – we consider we should have neutral institutions to keep peoples data and to use it just in peoples interests; 2) Jobs and Wages – we think it would be useful to find processes to put AI paying taxes and developing interest and activities for people.; 3) How to deal with Machine Learning and Human Enhancement?.		
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Richard	Graf	K-i-E.com		Extension for Human Centric AIArtificial intelligence today is cognitive computing. Human Centric AI demands that the inseparability of emotions, intuition and cognition be implemented in order to obtain artificial human intelligence (AhI). Human decisions are made by two systems (emotion and cognition system) in multiple serial, parallel and cyclic as well as interacting processes: affective decision, intuitive decision, conscious decision, rational decision, cognitive decision et cetera.For AhI it is necessary to develop the emotional logic and the process of inseparability. AhI would take into account that the emotion system is the origin and end of all thinking. New thinking with a conscious logic of emotions extends both human and artificial intelligence. Current AI focuses on		

				reproducing human intelligence, but people don’t always think or act intelligently. Rather than imitate human thought, AI should compensate for weaknesses in human decision-making to be truly valuable.Decisions precede every action and form the core of human existence. Modern life constantly requires us to make more decisions, faster. Evolution has equipped us with a decision-making system.• Professor at Goethe University, Frankfurt, Germany• Emotion Researcher since 27 Years • Author of the 2018 book, Die neue Entscheidungskultur (The New Decision Culture) • Founder, past-CEO of Simple Fact, A.G., Business Intelligence and Analytics.• Agile Coach and Project Manager for more the 200 projects• Diploma mathematician University of Regensburg, Germany			
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Charlotte	Sgier de Cerf	Federal Office of Communications OFCOM (OFCOM Email was not accepted..)	We strongly support that a mechanism will be put in place that enables all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis. It is important that the list of all stakeholders that have signed the Guidelines is made public in order to make their commitment and accountability visible.	With regard to Section 5 "Critical Concerns of the AI", we note that identification without consent is already addressed by Swiss legislation in particular by the Swiss Data Protection Act, as can normative & mass citizen scoring without consent. In general, it can be said that the current Swiss legislation to combat abuse is also applied in the online world, although for specific cases, e.g. when humans don’t know they interact with a machine or an AI identity, there is certainly still a need for adaptation. Major AI-enabled developments may also require in the future some specific legislative responses reflecting actual needs to protect, in the way of legally accompanying those developments.	With regard to Section 2 “Technical and Non-Technical Methods to achieve Trustworthy AI, Non-Technical Methods”, we stress the importance of education and awareness building to foster an ethical mind-set of not only the designers and developers, users (companies) but in particular also all the individuals of our society, using or not AI systems. Many people show today lack of trust in artificial intelligence because they do not understand its functionalities and its potential impact in their private and professional live. The education system on all levels, in particular in the field of vocational training, should ensure that ethics in AI becomes an integral part of education and training and incorporate these aspects into the expectations of responsible entrepreneurship.	In Switzerland, the ongoing revision of the Data Protection Act takes account of various constellations in which personal data are processed using algorithms. For example, an information and consultation obligation of the concerned person (data subject) is planned if a decision is made against him or her that is based exclusively on automated data processing and has legal effects or significant effects on him or her. With the right of access, the data subject should also be able to demand further information about the result, the conclusion and the effects of the decision. The project of law also contains measures regarding the profiling, which is often based on the use of algorithms. Furthermore, those responsible are to be obliged to prepare a data protection impact assessment if the processing could lead to a violation of the personality of the person concerned or of fundamental rights.	In our understanding, the debate on AI respecting fundamental rights, principles and values should be organized including all groups of stakeholders in order to promote a interdisciplinary debate on the decisions to be taken as well as to promote the informational self-determination of everybody. In this regard, we very much welcome the open and public stakeholders’ consultation process on the “Draft AI Ethics Guidelines” and the possibility to provide feedback. In our view, any policy framework on AI can only be developed, implemented and reviewed through international cooperation and multi-stakeholder dialogue across silos, in order to guarantee diversity of thought. Furthermore, the development of AI should not be left only to what can be done, but what should be done to promote the welfare of the society as a whole, with a human centric approach by protecting fundamental values, human rights and human dignity in the digital age.
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential

Maria	Sotiriou	GSMA		IntroductionThe GSMA supports the European Commission's endeavour to maximise the benefits of artificial intelligence (AI) while minimising the risks to individuals and communities, and appreciates the opportunity to comment on the Commission's new draft ethics guidelines. We support the view that the development and deployment of AI systems should respect fundamental human rights and applicable regulation, as well as principles and values ensuring an 'ethical purpose'. A growing number of GSMA members have already committed to responsible development of AI technologies. The general approach set out by the High-Level Expert Group on Artificial Intelligence (HLEG) in the consultation document is clear and well-considered. Only AI that is proportionate, trustworthy and robust has the potential to achieve mass market adoption. By emphasising trust, we are confident that EU technology providers will be strong players in an AI-driven world economy.We need only observe the global reaction to the General Data Protection Regulation (GDPR) to see how the exercise of 'soft power' based on a strong rules-based framework can indeed help to shape global markets and strengthen the EU economy. We see no reason why AI ethics should be any different, and the GSMA is strongly committed to working with the European Commission and multi-stakeholder groups to deliver on this ambition. As a general comment on the Draft Ethics Guidelines for Trustworthy AI, they should be focused as much as possible on AI, and the relationship between AI and human rights impact assessments and data protection impact assessments should be made more clear. The document repeats what is already set out elsewhere, and the reader would be given better guidance if it referred to more comprehensive documents on the general approach, such as the OECD Due Diligence Guidance for Responsible Business Conduct. Through such clarification, readers outside of Europe would also more readily accept the universality of the framework applied. With these changes, the document could focus more narrowly on good practice and safeguarding fundamental rights that are unique to AI, thereby increasing its impact on practical business.Our detailed comments on the guidelines are primarily intended to ensure consistency with existing regulatory frameworks and to advise where measures proposed by the HLEG are either disproportionate or technically unfeasible. In all cases, we have suggested alternative wording that should align the guidelines with industry best practice and ensure a clearer link with AI developments currently underway across the telecommunications ecosystem.We also observe that there are many beneficial applications of AI, for example in fraud prevention, mobile network optimisation and improved IT security, that should be encouraged rather than impeded by the guidelines. Our view is that AI technologies are implicitly useful tools whose application requires an appropriate level of human oversight and application of law, such as the GDPR and ePrivacy Regulation. Even though the guidelines are voluntary and of the 'soft law' nature, it is vital that they do not introduce new terminology or rules pertaining to areas that are well established in law — from human rights to privacy and	Section B, Chapter I: Respecting Fundamental Rights, Principles and Values - Ethical PurposeThe GSMA agrees with the fundamental rights, high-level principles and correlating values identified in the consultation document. At the same time, the terminology and content of the guidelines should be fully aligned with the legal terminology of such concepts as human rights and strive to avoid extended interpretation of these well-established areas of law.However, this response will focus more on the guidelines' potential implementation issues. Predictability on how to implement and monitor conformity with the guidelines is of paramount importance to mobile operators. In order to achieve the intended results, the GSMA proposes the following:Providers of AI technology should be empowered to contextualise and make adjustments to suit various use cases.AI is not legally defined in EU law. A definition should: 1. Exclude software systems based on traditional and determined algorithms which are clearly not based on AI. 2. Focus specifically on AI algorithms that require human supervision only when the purpose may constitute a risk to individuals' fundamental rights.3. Capture the fact that the AI algorithm takes decisions as a consequence of the application of advanced analytical techniques (machine learning, deep learning and natural language processing) in combination with automation advanced feedback loops to solve problems.4. Introduce a risk-based approach related to ethical issues: a. Benign AI algorithms should not be submitted to ethical scrutiny. For example, AI algorithms that act as recommendation engines for audio-visual content, speech recognition or translation should not be submitted for ethics scrutiny (only GDPR rules).b. Application of AI algorithms that may have legal or security effects on individuals should not be subject to ethics guidelines that duplicate or contradict their existing regulatory requirements via horizontal privacy/security laws.c. AI algorithms that may take lethal decisions, i.e. AI algorithms for weapons (LAWS) should be excluded.The GSMA considers the use of AI to fall into two broad categories: technology-focused AI and commercially focused AI. For the former, AI is used to assist with fault detection, predictive maintenance and network planning and optimisation, all of which enables operators to make more efficient use of their physical assets. Use cases in this category often do not involve processing of personal data and have little direct impact on the fundamental rights of individuals. AI is also used for commercial purposes such as pricing promotions, predictive care, smart retail and through the deployment of virtual assistants (such as Tobi, the Orange-Deutsche Telekom chatbot on the Djingo smart speaker and the Telefónica Aura virtual assistant) and more. The HLEG should establish at the outset that a one-size-fits-all approach is not appropriate, and that a determination of how ethics guidelines apply should be made on a case-by-case basis. For instance, evaluation of the ethical purpose will vary considerably between the use of AI in relation to virtual assistants and that of any of the areas mentioned in Chapter B.I.5 (critical concerns). The context in which AI is applied must always be kept in mind.• From Fundamental rights to	Section B, Chapter II: Realising Trustworthy AIThe implementation and realisation of trustworthy AI is critical for achieving the desired outcomes of the guidelines. Clarity that allows for predictability, understanding and policing of the guidelines will pave the way.• Implementation of the EU's Rights-Based Approach to Ethics [p.5]While the EU's focus is correctly based on the EU Treaties and Charter of Fundamental rights, the guidelines should ensure that ethics are considered in relation to how organisations comply with the law, or how they should act where the law does not specify or address the specific context. In other words, the guidelines should focus on how the EU's rights-based approach to ethics should be implemented. Ethical considerations and guidelines should not contradict legal requirements. Legal instruments, such as the The Universal Declaration of Human Rights (UDHR), the EU Charter of Fundamental Rights and the EU GDPR provide both terminology and basic requirements that will need to be reflected in the document. If EU legislation is changed, then such changes will need to be reflected in the document. It should also be noted that the same rights and protections should apply online as well as offline.It should be made clear that these guidelines do not call for new requirements in law, but instead encourage good practice in developing and applying AI while appropriately safeguarding fundamental rights. The guidelines should enable organisations to identify and weigh good and bad outcomes to determine the best course of action. As the law will generally reflect the needs of a range of stakeholders, the European rights-based approach provides the natural point of departure for guiding the principles and values that help to understand what 'good' and 'bad' practices may be.• Accountability [as part of 'Requirements of Trustworthy AI, p. 14]'Accountability' as described in the consultation document seems to focus only on redress and remediation. However, accountability goes beyond that. There are already multi-stakeholder efforts in place that encourage good practice, including multi-stakeholder efforts such as the Partnership on AI.• Data Governance [p.14]The GSMA emphasizes that the GDPR provides a robust and comprehensive framework for the processing of personal data involved in AI solutions. GDPR provisions, which tailor rules to the sensitivity of data and how it is used, and include data subject rights, are sufficient to address data governance and privacy concerns related to AI. Existing privacy principles are relevant here, e.g., the quality of the data sets (adequate, not excessive) and avoiding bias (fairness, impact assessment, privacy by design). As with the GDPR, this requirement should not become a disproportionate burden when implemented. The guidelines should ensure responsible approaches to data selection and training to avoid bias and discrimination. They should also encourage the necessary steps to ensure reliable AI performance.• Design for all [p.15]We are concerned that this principle lacks specificity: If something is intended to be prohibited or restricted, then the harms should be quite clearly articulated. There are unlawful forms of distinction (e.g., racial or gender-based discrimination) and lawful forms (e.g.,	Section B, Chapter III: Assessing Trustworthy AIThe GSMA considers the menu of potential assessment questions posed by the HLEG to be helpful to entities developing and using AI technologies in a manner consistent with human rights. It is important that any approach to assessment be flexible, reflecting the different types of AI solutions companies pursue, including those that have little direct impact on individual rights. • Ethics in autonomous systems and time-scales The GSMA and its members believe it to be instructive to look at ethics and intelligent autonomy based on time scales. For example, do we allow AI systems to take full control and decision autonomy below a certain time limit beyond the human capability (e.g., below 500 milliseconds)? What situations do we allow this to happen in? What about the time scales in which humans are able to intervene in automated decisions, or what if they do want to be in complete control? What about time scales where autonomous decisions may become reversible or iterative as more information becomes available? Just because there is sufficient time to reverse a decision, there may still be domains where AI should not to make such decisions without human oversight. However, below the 1-minute threshold, there may be many situations where it could be crucial to let the machine take control. This aspect has not been addressed at all, but it is very important. For example, network management functions such as beamforming in 5G networks — aimed at increasing spectrum efficiency — will require autonomous systems to make decisions in fractions of a second in order to ensure uninterrupted connectivity.Under existing data protection rules, it is already incumbent upon organisations to consider points of ethics and fairness in order to avoid harm. These require significant assessments, processes and record keeping. As mentioned before, any operational or practical guidelines developed in the context of AI should be aligned to the greatest extent possible in order to minimise duplication.• Technical robustness is most likely to lead to positive outcomes if underpinned by sustainable business modelsFor AI systems to be technically robust and reliable, sustainable business models are key. In the case of the mobile industry, intangible assets (data, insights, analysis, services) can be transferred from a mobile operator for use by a demand-side agency under mutually beneficial terms that enable an ongoing relationship between both partners. This aspect of sustainability allows for robust, repeatable and replicable use of mobile big data across different geographies and use cases, underpinned by secure funding that enables continuity in supply and analysis of the data. For more information please see https://www.gsma.com/betterfuture/resources/sustainable-business-models-report• Assessment ListThe GSMA supports the objective of this chapter to provide a practical checklist of questions to ensure the development of trustworthy/human-centred AI from an early stage of the development cycle. However, our concern with this section is that while the questions here are appropriate considerations, they lack the technical detail and specificity that would make them useful or practical for AI product developers or engineers, particularly within a small organisation. In general, there is too

data protection. We are concerned that the guidelines may be based on certain misconceptions of EU data protection law, especially the GDPR, and would suggest a reexamination of these interpretations before final publication. The main issues include excluding data protection and privacy from fundamental rights, incorrect terminology (PII vs. personal data), incorrect applicability of data subjects' rights (only erasure and portability mentioned as being interchangeable rights), and inaccurate reflection of the existing data protection obligations under EU law (e.g., legal grounds, transparency, automated decision-making).Section A: Rationale and Foresight of the Guidelines• Endorsement mechanism [p. 2]The introduction of a mechanism under which stakeholders will be able to 'formally endorse' (p. 2) the guidelines raises questions regarding its practicality: What are the consequences of an endorsement? Would this (fully or partly) replace self-regulatory initiatives such as codes of conduct or self-binding guidelines? Would signatories thereby fall under specific external governance/auditing? And would choosing not to sign these guidelines create a false impression that a stakeholder does not support ethical considerations regarding AI? Lastly, it appears difficult to achieve broad endorsement of the guidelines in the form of a 'take-it-or-leave-it' approach, where some guidance might be considered acceptable by those unwilling to accept the guidelines in total. While the intention to regularly update and evolve the guidelines by treating them as a 'living document' (page iv) is understandable, it might also lower stakeholders' willingness to endorse them formally. It is also important to note that governments and policymakers can likewise develop, deploy or use AI and thus also qualify as stakeholders.• Trustworthy AI [p. 2]We agree with the assessment that "no legal vacuum currently exists, as Europe already has regulation in place that applies to AI" (p. 2), not least due to the technology's cross-sectoral nature. While the guidelines are not intended "as a substitute to any form of policy-making or regulation" (p. 3), the aforementioned conclusion nevertheless must be taken into account for the HLEG's second deliverable, i.e., the AI Policy and Investment Recommendations, due in May 2019. In this context, we suggest a footnote clarifying that due to fast technological developments, the existing legal framework may need to be further developed and adapted to new requirements, such as with regard to cybersecurity and information security. When it comes to competition law, the authorities should be equipped with the necessary tools to intervene in cases of market abuse related to exclusive access to data and platforms and to address emerging issues such as algorithmic pricing.

Principles and Values [p.5]The proposed ethical principles represent a widely accepted approach to the development of AI, and they echo many of the principles released by GSMA member companies. We would also like to highlight that robust data governance mechanisms are essential for any business that focuses on data, including those that pursue AI solutions. The designation of a Data Protection Officer or Chief Privacy Officer, the adoption of strong policies and procedures, and a culture of compliance are precursors to the implementation of an ethical approach to AI.The GDPR should provide confidence that personal data is processed according to the regulation and that data subjects' rights are respected. In addition, public authorities should ensure that the population has a basic understanding of what AI entails by encouraging educational institutions to teach this topic. The HLEG proposes that 'informed consent' is a value needed to operationalise the principle of autonomy. In this context, the HLEG should note that current legislation does not require consent from individuals interacting with AI systems under all circumstances (cf. the principle of explicability). Both private and public sectors should be able to process personal data based on legal grounds other than consent, including when implementing AI technology. Consent is not necessarily a precondition for a human-centric or a privacy-friendly AI; instead, the balancing of interests required by the GDPR represents the proper approach. The GDPR allows for individual control in appropriate circumstances. For example, Article 22 allows data subjects to object to decision-making based solely on automated processing when the decisions produce legal effects concerning the data subject or similarly significantly affects him or her.While consent can be one solution to guarantee accountability and transparency towards users, it is not the only one. According to the GDPR, processing of personal data (including for the purpose of offering an AI-based service) is permissible when it is justified by one or more of six different legal bases (including consent), such as processing necessary for the performance of a contract or for legitimate interest. In addition, the principle of compatible further processing (Article 6(4) GDPR) allows companies to use personal data for purposes other than the initial basis without the need for an additional legal basis. Consent is thus not the sole value and solution to enhance explainability. Voluntary approaches such as a one-pager that explains in simple terms the purpose for which personal data is being collected can enhance transparency significantly. Therefore, the notion of informed consent is given too much prominence in these guidelines, creating a misleading perception that it is the only and best requirement to preserve autonomy and explainability.The GSMA supports the recommendation that consent of the data subject should be obtained in many circumstances, e.g., for the use of facial recognition technology, in line with the current privacy and data protection laws. At the same time, it would not be appropriate to provide notice and require consent when facial recognition technology is used in the course of a criminal investigation, for example.An important example for the mobile industry relates to differential pricing, age-restricted items). Is the intention to also limit the lawful distinctions? We ask the HLEG to provide additional clarity on lawful/unlawful differentiations in the context of the guidelines.• Governance of AI Autonomy (Human oversight) [p.15]Existing privacy principles may be helpful to consider here, such as GDPR Article 22 (referenced above). Companies should have the flexibility to decide how to best operationalise this requirement in a proportionate manner. In our view, it is not necessary to guarantee human control at all levels (e.g., where AI is deployed deep in the network for fault detection). Human control may in some cases only be necessary in setting the outcomes, whereas in other cases, where there is a significant impact on individuals, human control is essential. The 'conservative approach' and 'human oversight' principles should help with this objective. Human control and a stop-button failsafe may not be necessary precautions for all types of self-learning AI approaches. Indeed, if we mandate these types of control, even for AI that is deployed deep in our networks and has no human interaction or customer-facing element, we could deprive AI of its greatest potential: to solve problems (like cancer treatment, reversing global warming etc.) that humans have not been able to. We request that the HLEG give more detailed thought to some of these subsidiary questions before including human control and/or stop buttons as a requirement in this section of the guidelines. A contextual understanding of AI, where different use cases are permitted with differing levels of human control, appears to us to be the optimum outcome.• Respect for (& Enhancement of) Human Autonomy [p.16]. The notion that an AI system would result in abuse should lead to an obligation to re-assess the requirements for Trustworthy AI as described in chapter III.We note that AI services are already deployed successfully in recommendation systems, as used in e-commerce sites and media consumption, as well as services such as search engines. The suggestion to allow the user to specify preferences and limits for system intervention is something we believe is covered already under the GDPR as part of the consent requirements as well as ePrivacy regulations. In practice, as AI solutions become more sophisticated, we have concerns that it will be difficult to provide fine user controls that influence the outcomes of AI solutions. We think it would be better to focus this section on the following ethical practices:• That AI not be designed to deceive users, for example by the creation of virtual users, customers, reviews, etc., which manipulate the behaviour of users based on peer views;• That AI be designed to be fair, e.g., not to suppress bad reviews of a product or service, or bias results in such a way so as to purely maximise profits at the expense of customer requirements or interests; and• That bias (racial, gender, age, etc.) be knowingly engineered out of AI systems both when the AI system is originally designed and as learning adapts.Earlier we noted that there are legitimate applications for AI such as fraud prevention and network optimisation that will benefit mobile users greatly. We think it would negatively affect users if there was the ability to opt out of

much repetition of the early sections of the guidelines and not enough effort to provide a clear structure for AI developers that can be easily understood in a variety of languages and for different AI use cases. The list is currently too broad and covers many requirements that are not specific to AI. The work would benefit from being even more specific about new challenges. The next steps could also benefit from including some less sensitive use cases of AI. The current four are all use cases where everyone agrees on the high risks towards safety, trust and ethics. It would be good to see how the assessment list would look for less risky use cases, such as customer service applications, marketing or similar.

mobile network optimisation, which customers have a right to expect as part of network service delivery. This will be a key area for the application of AI. Again, the guidelines should not create new rules and interpretations of the existing legislation (in this case regarding consent).• The Principle of Beneficence: "Do Good" [p.8]The GSMA supports the principle of explicability, while noting that some uses of AI technology will require explanation to data subjects and the public generally, and others will only require a business to explain elements of its technology to a regulator or expert body. Determinations regarding the level of explicability and transparency should be made according to the level of risk to individuals presented by an AI solution. In keeping with this approach, a business's use of internal review boards or consultation with independent experts should be considered good practice. • The Principle of Non maleficence: "Do no Harm" [p.9]Reference should be made to the 'risk-based approach' that underpins the GDPR, e.g., through the use of privacy by design, data protection impact assessments and the evaluation of data breaches to determine whether data subjects need to be notified, etc. Some of these concepts could easily be grafted onto the AI ethics guidelines to avoid reinventing the wheel. Some work has been done around the classification of harm in the privacy context, which could be leveraged for the AI context, although harms that would not directly affect individuals would need to be considered. This approach would also recognize that responsibility and flexibility may be more effective than regulation. In addition to the privacy principles, one could add that AI should contribute positively to the UN Sustainable Development Goals.We refer the HLEG to the UN Guiding Principles Reporting Framework, which enshrines the duty of states to protect human rights and of corporate entities to respect human rights. The second sentence of the section on 'Do no Harm' should therefore be amended to: "AI implementations should respect the dignity, integrity, liberty, privacy safety and security of human beings in society and at work."• Principle of Justice [p. 10]Instead of stressing "that AI systems must provide users with effective redress if harm occurs," the guidelines should emphasize that ultimately humans are responsible. Operators of AI should know and make clear who is responsible for which AI system or feature.• The Principle of Explicability: "Operate transparently" [p. 10]The guidelines need to be clear if explicability should be required for all AI systems or only for those that can potentially have a negative impact on their users if the wrong decision is taken. The principle of explicability should be proportionate to the level of harm that the AI system can cause.Critical Concerns raised by AIIt is helpful to consider concerns, and we understand how different points of view can coexist. Here are some initial thoughts of the GSMA members:• Identification without consentIt would be worthy to differentiate between 'identification' as a goal and as a side-effect. AI allows for easier identification of individuals (e.g. via facial recognition). The emphasis should thus be on not abusing this functionality. Use of AI-enabled identification and surveillance processes should follow current legal practices. The such beneficial AI services. • Robustness — Resilience to Attack [p. 17] The requirements described in the guidelines regarding resilience and robustness apply to AI systems as well as to any ICT system (e.g., IoT systems). Having said that, the guidelines would benefit from further consideration of the precautions that can be taken to raise the security level of AI systems. The highest security requirements should apply in AI development and application. All security features such as notification of security vulnerabilities, emergency stop buttons or security updates should be aimed towards a clear attribution of responsibility. Besides the risk of weak spots being exploited by hackers, the self-learning capabilities of corrupted AI systems raise the risk of damage exponentially. For security in the development and application of AI, this specifically means that ensuring IT security is a key requirement for product safety of AI applications or products that implement AI applications. This correlation must always be considered by developers and industrial users (i.e., security by design). Mandatory risk assessments analogous to the data protection impact assessment of the GDPR could contribute to highly sensitive AI applications, as in healthcare. The current regulatory focus on operators of critical IT infrastructures as in the ICT, healthcare or energy sectors, is no longer sufficient because critical issues arise increasingly on an ad-hoc basis. This would for example be the case in a multitude of connected, self-driving vehicles. • Respect for Privacy [p.17]The GDPR is principles-based, and this enables it to accommodate new technologies including AI. The GDPR is also based on the identification of risk of harm and on the concept of accountability so that organisations are encouraged to adopt technological and operational measures to control risk, including privacy by design, data-privacy impact assessments, the appointment of a Data Protection Officer, good record-keeping and being able to demonstrate compliance. To the extent that an AI deployment makes use of personal data, it is already regulated by the GDPR. The draft guidelines should therefore acknowledge this explicitly and avoid the duplication of requirements, which could cause uncertainty. Most of the harms discussed in the draft guidelines are in fact privacy harms. Privacy and data protection are separate issues. Some of the principles of the GDPR would have to be extended to accommodate a range of harms, such as harm to groups of individuals, harm to society, harm to the environment, and organisational responses to the GDPR would have to be adapted accordingly. However, the core mechanisms of basing requirements on flexible principles and the identification of risk of harm are already in the GDPR and could be extended to or replicated in the AI context. The requirements proposed in the draft guidelines are also very similar to privacy requirements (e.g., transparency, explicability, right to know when data is being collected) and do not need to be restated or reinvented just because AI is processing the personal data. The GDPR also recognizes that pseudonymisation and encryption, which facilitate beneficial uses of data while reducing the risk of harm to individuals' privacy are good practices. We therefore suggest these safeguards as

emphasis here should be on reliable anonymisation/de-identification methods and adhering to the GDPR. • Covert AI systems [p. 11]We generally support the recommendation that people should always know whether they are interacting with a human being or a machine. However, companies should have the flexibility to implement this requirement in the best way. • Information on process, purpose and methodology of the scoring [p.12]For complex systems, the GSMA has doubts that it will be practical to understand the logic developed by AI in order to explain to customers or even IT specialists how a decision is made by AI. In practice, decision-making in complex systems that do not use AI can be similarly difficult to understand, so the GSMA does not consider this to be a unique attribute of AI systems, although the self-learning nature of AI can be a barrier at scale. The GSMA would advocate that there is a focus on the learning process for AI systems, including strong human oversight on matters such as data set selection, target-setting and verification of results to ensure there is robustness and fairness in the automated processing implemented by AI. Additionally, this situation needs to be aligned with the privacy right to know the logic behind automated processing.• Tensions between individuals and societyThe GSMA agrees with the remark in Chapter I.4 (p.8), that "tensions may arise between the principles when considered from the point of view of an individual compared with the point of view of society, and vice versa." There will be AI solutions that are good for society (e.g. for the environment, disaster relief, healthcare, optimisation of public transportation, national security, immigration control) but that individuals may not perceive as bringing immediate personal benefit. For instance, societal changes brought by an increased use of AI or production/business processes may lead to temporary unemployment, which may raise a negative perception of the use of AI in the population. Public authorities need to consider how to manage such societal changes, in partnership with the private sector and civil society. Regulators should be open to innovation and innovative AI solutions, and only intervene when the legal and human rights of individuals are at risk. Reliance on the EU Treaties and Charter, as well as existing and new case law where the aforementioned conflict of interest has been addressed, shall become significant.It is of the utmost importance that private operators should not be expected to make a determination on where the correct balance lies between fundamental rights. Such determinations should only be made by actors with a clear public mandate to decide where the appropriate balance should lie — usually judicial authorities or elected officials.As it is impossible to foresee all intended or unintended consequences, even with technical and nontechnical methods in place, as discussed in Chapter II.2, it would be good to recommend establishing an AI ethical committee at the governmental level. This could work like the ethic committees established in each Member State in the area of clinical trials according to Directive 2001/20/EC. The purpose would be to provide guidance and create debate about new uses of AI that impact people and societies at large.

additional technical methods in the guidelines. Pseudonymisation has an advantage vis à vis anonymous data, namely that the necessary 'identifiers' remain intact for big data applications, to be able to merge large amounts of data from various sources. The technique thereby eliminates the direct link between the data and the data subject, while the pseudonym used as an identifier allows to repeatedly merge data from different sources over a period of time. This is a key requirement for valuable data-driven services, also in the field of AI. Private companies can thus play a role through the introduction of technical measures which reduce the need for difficult tradeoffs. In the context of the ePrivacy reform, for example, pseudonymisation could be deployed to ensure that electronic communications data can be used without impinging on fundamental rights. In addition, the challenges related to AI and the principles of data minimisation and purpose limitation should be emphasised. AI will also inherently challenge the principle of transparency. If absolute transparency is a condition, it will rule out deep learning.• Transparency [p.18]Again, much can be learned from EU data protection law, where processing of data is intended for a specified purpose (purpose limitation). Although "informed consent should be sought" sounds appealing, it is not always the most appropriate way to protect people, and it is extremely difficult in practice. The guidelines should not reinvent the wheel vis à vis existing legislation. Overreliance on informed consent could lead to people agreeing to everything or, on the other hand, create an insurmountable barrier to the good outcomes that AI could achieve.However, if a consumer's data is being used to make decisions about that person through the use of AI, they should be informed that this is happening. Where personal data is being used, an individual already has the right to opt out under the GDPR, so there is no need for an additional requirement.The context of AI use is very relevant here. If AI is used within a communications network to improve energy efficiency or routing, there should be no need for a consumer to have a right to opt in or out of such use.Regarding accountability, the guidelines should be clear on the kind of accountability intended. If this merely implies that organisations will be held liable under the law, then the issues to be explored are: Who is liable and under what circumstances? But, most important, the guidelines are not the right legal instrument to determine liability. If they are a call to organisations to hold themselves to a high ethical standard regardless of the law, then this should be the starting point for a chapter about how to enable ethical decision-making in practice.The section talks about explaining how the system makes decisions, rather than explaining the system's decision. This should be more clear. Neural networks are difficult to explain, but their decisions can, to some extent, be explained. Technical methods• Ethics & Rule of law by design (X-by-design) [p.19]A clear distinction should be made between safety-critical or ethics-critical systems and non-critical systems when demanding failsafe shutdown mechanisms and robustness vs adversarial examples. Demanding this for all AI systems (e.g., a music recommender) does not seem practical and will hinder

					innovation and commercial adoption of AI. • Architectures for Trustworthy AI [p.19]Integrating ethical goals and requirements at sense level for adaptive and learning systems is certainly the preferred way, but not the only way. Unwanted actions could also be filtered out. The latter is probably much easier. • Testing & validating [p.20]Testing within predictable bounds should be made offline, before deployment. It is even more important to monitor in the real world continuously.		
Egle	Gudelyte Harvey	Telia Company	Telia Company welcomes the Commission's proactiveness in addressing such an important topic and greets its efforts to include a wide variety of stakeholders in order to capture the essence of the benefits and challenges presented by AI. At its core, AI is a part is science and is not "magic", therefore it is important to approach it in a systematic manner. It is positive that the Commission has high ambition as well as encourages Member States to adopt AI strategies. Europe needs to get AI development right and one of the most important aspects is to build human-centric and Trusted AI. We support the view that the development and deployment of AI systems should respect fundamental human rights, existing legislation, principles and values ensuring an 'ethical purpose'.Telia Company's has committed to work actively towards the United Nation's Sustainable Development Goals. We embrace the value and opportunity of AI as an accelerator for realizing the 2030 Agenda for Sustainable Development. We aspire to integrate sustainable, responsible business practices into all parts of business and strategy to harness AI for good. We would like to contribute to the notion that AI can extend and complement human abilities rather than lessen or restrict them.Developing EU draft AI Ethics Guidelines (Guidelines) is a good start for the EU to set the right tone and ambition on the subject matter. At the same time, the actual draft would benefit from	The HLEG proposes that 'informed consent' is a value needed to operationalize the principle of autonomy in practice'. Based on the current legislation it is therefore not correct to state that a consent from individuals interacting with AI systems always shall be a requirement, cf. the Principle of Explicability, "Operate transparently". Both private and public sector should be able to process personal data based on other legal grounds than consent, also when implementing AI-technology. Consent is not necessarily a precondition for a human-centric or a privacy friendly AI. Article 22 in GDPR, which regulates automated individual decision-making, only requires consent when the decisions has legal effects concerning the data subject or similarly significantly affects him or her. Face recognition is already used as a crime fighting measure, and it will be meaningless to speak about consent in such contexts. Another example for the mobile industry relates to mobile network optimisation, as part of network service delivery and if legislation allows that consent is not necessary, then that is also the correct approach in AI environment. The guidelines shall not create different rules and interpretations of the existing legislation (in this case the usage of consent).Principle of Explicability is very important and shall be preserved. However we suggest proportionality and the risk based approach shall be implemented and applied to not to	Implementation of the EU's Rights' Based Approach to EthicsEthical considerations and guidelines cannot contradict legal requirements. Legal instruments, such as the Universal Declaration of Human Rights (UDHR), the EU Charter of fundamental rights, GDPR, etc., provide both terminology and basic requirements, which will need to be reflected in the document. Data Governance. The text outlines only principles for the data, which is input to AI, either in the training or use phase. AI algorithms shall also be developed and governed not to produce malicious data as an output, hence we suggest to address that. This requirement could also be considered to be part of requirement on "Safety". Design for all. "Systems should be designed in a way that allows all citizens to use the products or services, regardless of their age (…)" – we understand the intention and the equality aspect here, but positive discrimination exist already now under legislation, i.e. there are numerous products and services where clear age restrictions are present (age to be able to hold drivers license, sale of alcohol for minors, etc).Governance of AI Autonomy (Human oversight). Existing data protection legislation shall be applicable here as well, including while dealing with profiling or automated decision making. Stakeholders should have the flexibility to choose the most efficient way how to best operationalise this requirement in a proportionate manner and in line with the laws.Robustness.	Privacy.Personal data should be anonymized or pseudonymized whenever possible, before they are processed in an AI-context. However in the Guidelines focus has been to underline the need for consent from the data subjects, which is not in line with the current legislation which provides different legal grounds for processing of personal data.	On the definition of AI. According to the presented definition of AI it is still difficult to distinguish what are specific minimal requirements for software or hardware to be understood as AI (to avoid unintentionally capturing simple software and hardware).We would like to propose to add a notion that "Artificial intelligence (AI) refers to systems, parts of systems and/or technologies ...".We think that the base and the essence of AI is software, there is no possibility to have AI without a software, hardware is just a tool to realize some functions of AI in real (non virtual) life, therefore system (software plus hardware) can be understood as AI driven hardware.We would also suggest to remove the statement "(according to pre-defined parameters)" – it has a very classic "Business Process Automation" tone to it. The whole point with AI is that there is no need to pre-define every single parameter, the whole notion of AI is to use computation to learn patterns in data rather than having humans deciding all possible rules/parameters needed to take a particular aspect. We suggest either to remove the statement in parenthesis or reformulate it to include that parameters are learned from data rather than being pre-defined.The learning statement is not all encompassing, since learning is explained as only the part where an AI has acted in an environment, measures the response and optimizes accordingly (e.g. reinforcement learning). But there is also another side to it, namely

			scrutiny of the EC units responsible for already adopted legislation, such as GDPR. Ethical considerations and guidelines cannot contradict legal requirements and create different terminology, rules or interpretations of existing legislation. This is in particular evident where the use of consent is implied throughout the document. Even though the guidelines are voluntary and of the "soft law" nature, it is vital that they shall not introduce different terminology or rules when dealing with the areas that are well established in hard laws – from human rights to privacy and data protection. Endorsement mechanism. It is also important to note that governments and policy-makers can likewise develop, deploy or use AI and thus likewise shall qualify as stakeholders. Endorsement as such is a positive initiative but the practicalities in terms of what exactly would that entail remains to be seen.Scope of the Guidelines. We strongly support the text in the Guidelines which state that while the Guidelines' scope covers AI applications in general, it should be borne in mind that different situations raise different challenges and that context is important. We would further advocate for risk based approach. "One size fits all" approach for all AI applications will not be appropriate since different applications render different levels of risks and consequences.	absolutely all AI systems, but on those that can potentially have negative impact on their users. In case of deep learning algorithms it is difficult or practically impossible to explain with absolute precision how a certain input to these algorithms create a certain output. The relation is based on the learning algorithm and the data given to train it. It would be appreciated if HLEG could propose a model answer(s) for the deep learning case.Critical concerns. We suggest to add new concern "AI systems vulnerable for malicious external interference". Poorly designed or tested AI systems, which can be influenced by external means to change completely their output and reaction from the designed targets. Example – chatbots turning to sexist racists with the use of biased input."Covert AI systems" – the context in which an AI is employed as an important distinction to be made. This is fundamental and extremely important aspect to be considered. A lot of the AI telco use cases would operate within a context where personal data is not used and would not impact individuals.Identification without consent. All identification and processing of personal data by means of AI obviously requires a valid legal ground under the GDPR. Clear legal understanding of what legal grounds can be employed in the context of AI needs to be developed on EU level. Tensions between individuals and society. It is of the utmost importance that it should not be left to private operators to make a determination on where the correct balance lies between fundamental rights.	Although testing is understood typically to be part of development phase, we would like it to be mentioned separately: "Trustworthy AI requires that algorithms are secure, reliable as well as robust enough to deal with errors or inconsistencies during the design, development, testing, execution, deployment and use phase of the AI system, and to adequately cope with erroneous outcomes."Privacy. Although the guidelines are universal and general and therefore cannot go too much into details, the legal and ethical issues relating to privacy should be addressed in a more concrete and tangible manner. AI and the challenges related to the principles of data minimization and purpose limitation should be emphasized. AI will also, by nature, challenge the principle of transparency. If absolute transparency is a condition, it will rule out deep learning. Technical Methods to achieve Trustworthy AI. We suggest to add the human governance aspect to these methods. AI systems should be designed in a way that human operators can always monitor, control and shut down the AI system even under irregular conditions.Key guidance in realizing trustworthy AI. We suggest to add documentation and logging of key decisions, design, data sources, training plans, test plans and results, operations instructions etc. in AI development and use (some of these are mentioned under the Chapter III lists, but shall be better placed as guidance here).		that in order to train the AI in the first place, it is able to learn based purely on evaluating (often very large sets of) data using significant computing power. There is where the learning aspect is used mostly today, and it is what currently most distinguishes AI from "classic/non-AI" systems.Other comments.The guidelines currently contain misleading information regarding data protection laws, especially the GDPR, and require the revision in that area before publishing. The main issues include excluding data protection and privacy from fundamental rights, incorrect applicability of data subjects' rights, and overall lack of correct reflection of the existing data protection obligations under EU law (legal grounds, transparency, automated decision-making).The Guidelines should use the language from existing legislation. Inconsistencies would lead to confusion. Legal instruments, such as the Universal Declaration of Human Rights (UDHR), the EU Charter of fundamental rights, the EU GDPR, etc., provide both terminology and basic requirements which will need to be reflected in the Guidelines. If such legislation will be changed, then such changes will need to be reflected in the Guidelines. It should also be noted that the same rights and protections should apply on-line as off-line.
Anonymous	Anonymous	Anonymous	As can be observed already by today, the development, rollout and technological adaption of Artificial Intelligence (AI) by civil, governmental and corporational bodies will lead to fundamental changes in the entire spectrum of our European society. Yet, to understand which laws and regulations are needed to gain a maximum benefit from AI technology for society, as well as to guarantee that these benefits are fairly shared between all of its members, the process of juridical frameworking and regulation of and around AI, including a basic ethical guideline of how to deal with AI, has to be designed from the start as ongoing and transparent and thus able, to adapt to developments which become first visible after the wide-spread introduction of the technology. Constant evaluation on current changes caused by AI is therefore needed, e.g. by setting up the legislative premises to equip the European Parliament with the necessary power to debate and decide on regulations concerning governmental and corporation-related use of and investment into AI.	regarding passage: Trustworthy AI (page 1): AI is thus not an end to itself, but can rather, when responsibility is taken by citizens, governments and corporations alike, be a means to increase individual and societal well-being. Issues of diversity and inclusion (with regards to training data and the ends to which AI serves), should be a focal point, when designing and introducing changes based on AI. Benefits of AI have to be shared equally among all members of society, therefore the question on distributive justice should be asked in the light of everybody participating in benefits gained by AI. Scope of th Guidelines (page 3): Likewise, different opportunities and challenges arise from AI systems in the context of business-to-consumer, business-to-business, public-to-citizen and citizen-to-citizen relationships, or.. (no more suggestions for changes in this sentence). a comment to the graphic on page 4: "Framework for Trustworthy AI". I miss two central points: where are checks and balances and where is the role of the European Parliament and the European Commission in this graphic, thus in the process? I. Respecting Fundamental Rights.. (page 5): Fundamental rights cannot only inspire new and specific regulatory instruments, they should be the guidelines for AI systems' development..(no more suggestions for this sentence).			I very much miss the mentioning of the responsibilites of and regulations for companies regarding AI, throughout the enitre document! This is only included for governments and should definitely be changed. I furthermore wonder, why non-European companies like Google and IBM were part ot the HLEG. As the past and present shows, e.g. in conflicts about tax evasion and abuse of market leader positions, European and especially non-EU companies already try to avoid regulations valid in the EU. I furthermore want to criticize the entire stakeholder process on AI as weakly transparent and too short. Comparing it to the recently introduced committee on AI by the German Bundestag, set up to work for two entire years, 4 and now 6 weeks seems way to short. Furthermore, the timing of the stakeholder process over christmas and new-years brake, when people do not (and have the right to not) care as much about political decisions as during the rest of the year puts the entire process under the shabby light of vested interests by the consulting and involved companies and institutions of the HLEG. Acts like this will increase peoples' critical position about intransparency of decisions and processes by the EU as a whole.

			2. From Fundamental rights to Principles and Values ...to make an educated decision as to whether or not they will develop, use, or invest in an AI system at experimental or commercial stages, including the right to decide to not at all use a governmental or commercial system based on AI, without suffering under disadvantages regarding the persons human rights and needs (no more suggestions for this sentence) 3. Fundamental Rights of Human Beings 3.2 Freedom of the individual Every person must have the right to decide to not at all use a governmental or commercial system based on AI, without suffering under disadvantages regarding the persons fundamental human rights and needs (no more suggestions for this sentence) 3.5 Citizens rights At the same time, citizens must enjoy a guaranteed right to be informed of any automated treatment of their data by government bodies or companies, and systematically be offered to express opt out. Citizens should never be subject to systematic scoring, neither by the government nor by private companies or similar institutionalized organizations. Citizens should enjoy a right to vote and to be elected in democratic assemblies and institutions. To safeguard citizens' vote, governments and private companies must take every possible measure to ensure full security of democratic processes.				
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Anonymous	Anonymous	Anonymous	I am a bit disappointed by the text. It often constitutes a verbose instance of political wish lists, too often without AI-specific messages. It also seems driven by concepts from the humanities exhibiting sometimes a limited understanding of technological and economical realities.It believes in ethical AI by design, which - like the concept of "ethical computers", if taken too seriously, and ignoring a probable theoretical impossibility - would cripple most real AI-systems. The concept of trustworthy AI is not so obvious to implement. I would not trust AIs forced to realize political agendas, nor those with a shallow, narrow view of the world, incapable of reasoning and deeper analysis, which is however characteristic of the currently prominent ML-algorithms.Given the strong demands formulated in chapter II, I see a strong risk for stifling innovation.Especially if one does not just indicate rough guidelines, to be interpreted reasonably by humans in changing contexts, but if one tries to really enforce specific rules potentially in direct conflict with what makes AI fascinating and game changing.It may therefore be too early to implement many AI-specific guidelines beyond those in place for human-driven tasks or data management. This would hinder the development of AIs and also raise expectations which in my eyes are impossible to meet. So by emphasizing these	It also sometimes confuses AI with data-crunching/mining. There are already - partly controversial - rules related to data collection and processing imposed upon companies and institutions, which would of course also apply to companies using among others AI technologies. It is then the question whether in AI applications, there should be higher requrements as in human-driven processing - I would say, in general, no. It is a technological problem partly linked to the sophistication of the evolving programs and systems how to realize similar or complementary outcomes (maybe humans are better in one kind of decisions, AIs better in another one).There is furthermore the implicit idea that there are clearly defined absolute standards. Human history and global diversity show however that even for accepted standards, given their vagueness, the interpretations can vary considerably. Consequently, also for commercial reasons, we should make sure that our systems are flexible enough to allow for enough synchronic and diachronic accommodations. The problem is also less individual AIs gathering and processing data, but their automatic sharing of knowledge with all the other AIs in companies or institutions which creates risks.A problem I see is that existing regulations may not take into account the specific needs of AI (see GDPR). It should not be just about inventing	Among the requirements in chapter II, I would support as AI-specific and relevant:1, 2? (in the sense of ensuring the use of available background knowledge, not in the sense of replacing a data bias by a political one), 4, 7, 8, 9, 10. The other issues may be worthwhile to some extent but their handling is not an AI topic, resp. follows from other points. Ethics by design. Ensuring ethics by design is an impossible demand if understood in an absolute sense. There may be parameters to set, but this is user-driven information, possibly required by the company or state. Security by design is reasonable, privacy by design depends on how far this is realizable in humans. The problem is less about collecting and processing data than about sharing them.Architecture is relevant insofar as each system should have a reasoning and communicating entity allowing for humans to interact meaningfully with the system and to get explanation. Also there should be safguards against manipulability. An isolated deep learning system typically does not match these demands. Testing and validation is important, but in reality there are always tradeoffs. We have to see that a strict application of this methodology to humans may well have negative effects. Similarly for AIs. So let's avoid AIism.Traceability, autitability, and explanation are certainly relevant.At the	Concerning the requirements, I would add the need of having "Sufficient Intelligence", i.e. rational methods for reasoning and planning, access to verified knowledge, communication ability, and the best available ML-algorithms in line with the available resources.Another important point missing is the call for AIs able to understand humans and interact in a natural manner with them, sufficiently in line with human cognition. This is for me the most relevant meaning of human-centric AI.I would think less about imposing requirements than about dropping some of them, especially those linked to non-AI-specific themes (e.g. those about data, digitization) or primarily driven by specific political agendas without near universal acceptance, or culturally biased. To summarize, while I agree with the importance of ethical considerations, especially as a technological challenge, I would subscribe at most to half of the demands made here. The other ones may well be detrimental to AI, whether human-centric or not.	The comments reflect some quick personal thoughts, neither those of my employer, nor necessarily those of CLAIRE,which I support in a constructively critical way.

criteria we risk, after possibly increasing trust in AI in the near term, to seriously undermine it in the long term because this will be neither commercially sustainable, nor feasible.The goal should not be to impose particular ethical viewpoints but to ask for algorithms which are open for a deeper analysis necessary for many instances of human-centric computing.The demands should in general not be higher than those for humans, companies, or institutions.They should also be streamlined. It makes no sense to just repeat all whole discussions on data management and data mining in a digital world. It is enough to refer to them and to focus on specific AI-themes, going beyond this. Similarly for other areas, e.g. concerning social scoring, which is not at all dependent on AI. So while one may well support certain ideas, they should be handled separately, and certainly not at a level where technological standards are set.

new possibly counterproductive rules, but also about questioning existing rules to prepare for a new world with (also) beneficial AIs. Ethical rules for AI should focus on politically neutral ethical demands and not subscribe specific left- or right-wing agendas. Otherwise this whole initiative will be brandmarked as a political initiative and go nowhere.Assuming that there is a generally accepted idea of a common good to be necessarily targeted by AI-systems is incredibly naive and does not in any way reflect economical or political reality. But technology standards are not the place for socio-political utopias.What is also missing from the document is an explicit consideration of uncertainty which means that the decisions about what to do are necessarily imperfect even if we assume - unrealistically - that there are clear utilities. The main goal should be to promote AI-systems with a sufficiently deep understanding of the world with strong capacities of practical and normative reasoning, able to evolve. The rules may actually be less constraining for these systems than for the much more common, essentially statistical algorithms on speed.The role of a technology is not to protect "the democratic process and the rule of law". This is much too far-fetched. Within a society AIs has to observe the given rules, but these rules may well have to be adapted to optimize a world with AIs. If taken at face value, with the above principle there would have been no internet and no public cryptography. Like any relevant civilization-pushing technology, there are good and evil applications, noting that there is not always agreement about what's good, or not.A general right to refuse AI-services looks completely non-sensical. Similarly, certain amounts of (accepted) subordination are necessary in a human-only-world, this has to be accepted also in a mixed world. It is a question of the quality of the services, often to be determined by market forces. There should be no a priori discrimination toewards AI-systems based on some naive views of presumable human needs. Among the ethical principles from II.4, the only one which makes full sense to me is explicability. Humans can rightfully expect explanations of decisions and judgments from other humans or institutions, and this should also be possible w.r.t. AIs.Identification without consent is not the problem, it is the global sharing of this information with dubious forces, sometimes including the state, which is critical. Similarly, the problem with personal assistants is not that they come to know the family, but that this information ends up in corporate databases beyond control of the consumer. Let's turn towards the critical concerns. So, 5.1 is not an AI issue. Concerning 5.2, I see absolutely no problem with covert AI systems. 5.3 concerns mass surveillance, linked to AI but this is more an issue of digitization and politics considered totalitarian in Europe. That we use inteligent weapons to defend Europe and its values against evil forces, and to minimize damage to civilians, seems not only acceptable but even an obligation. I would completely reject 5.4. In the science fiction scenario addressed in 5.5 , I see it as a highly interesting approach to build artificial consciousness, and I see a necessity to have artificial agents doing moral reasoning by themselves.

current stage, regulation specifically directed at AI seems counterproductive and hindering scientific and technological progress. Some rules may however be derivable from expectations we have towards humans.Standardization, as commonly understood, seems absurd at this stage of the AI technology. This might actually prevent longer-term solutions. There could however be demands e.g. concerning a human-like cognitive architecture which would make sense.I see not too many needs for accountability beyond those in place anyway. Codes of conduct can also be stifling. Education is important so that people understand what is going on, what expectations one may have, and what some may consider risks. For developers, education about security and privacy issues would be most relevant. Involving stakeholders is good to hear other views, but technological progress in our risk-averse society is impossible if one insists on a consensus here.Diversity is a popular issue in some circles, but I think that we should more worry about qualifications, education, and motivation. The application and societal rules determine which needs are to be satisfied, not abstract demands on the composition of a team.

Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Nicolas	Moës	Acting in private capacity.		Section 5. Critical concerns raised by AI; Subsection 5.5. Potential longer-term concerns.I would argue in favour of keeping and even expanding on these controversial issues. I do not argue on the object-level plausibility of these longer-term concerns realising themselves (as a matter of fact I am skeptical for some of them), but I have 4 higher-order arguments supporting my case. 1) Citizens' freedom to decide and the expert group's intentionality:First, this expert group's objective is to generate ethical guidelines. These will guide public opinion and, perhaps, policies, business CSR strategies, adjudications, ... but they will not constrain them. As an inspiration, these ethical guidelines should empower citizens to consider a broad range of potential issues rather than to paternalistically decide, by foregoing it, that an issue is improper for citizens to discuss. The guidelines indeed still allow citizens, policymakers and entrepreneurs to disregard these longer-term issues if it turns out they do not ever materialise. Duly acknowledging the existence of these longer-term concerns among the expert community and the population and even explaining them would therefore be the responsible course of action, in line with European values of citizens' empowerment and freedom. 2) Precautionary principle and the expert group's legitimacy:Second, as an expert group discussing ethics rather laws and legislation, it seems necessary to start from the world we would like to live in, rather than the current situation and these political economic considerations. Through that lens, precaution is much more prominent. This is especially true in Europe where a history of technological accidents and scandals in the 20th century have made the precautionary principle the foundation of techno-scientific policy, jointly upheld by European nations. In this case, precaution requires at the very least acknowledgment and explanation of these longer-term issues. Not exercising precaution by dismissing these longer-term issues as mere footnotes would be discrediting the expert group's efforts to express and embody European, human-centric values on behalf of citizens. This would undermine the perceived legitimacy of all the other guidelines from this expert group. 3) The EU as global champion of ethics and the expert group's credibility:Third, more practically, some of these "highly controversial" issues have apparently already been accepted and supported by democratic legislatures, with regards to Artificial General Intelligence (AGI) at least. The State of California has decreed guiding values for the development of artificial intelligence, which are principles tailored for ensuring the beneficial development of AGI (Assembly Concurrent Resolution No. 215, September 2018). If the State of California, whose GDP depends on the tech industry's freedom from overregulation, decides to endorse precautionary principles on AGI in official resolutions, the European Union's ethical guidelines probably also ought to contain a thorough explanation of concerns in that area. The stake here is not only the		Disclaimer: the views presented here do not necessarily represent the views of my partners and clients.

credibility of the expert group, but also the credibility of the European Union as a champion of ethics in AI globally. A genuine champion would not sweep controversial issues under the rug - especially not if these are fully acknowledged and acted upon by other states.4) Population's need for guidance and the expert group's duty:Finally, on its own, the presence of controverse among your expert group is sufficient evidence in favour of including the highly controversial topics. If there is disagreement on ethical guidelines, this disagreement ought to be flagged and the controversial issues explained. Indeed, even if just a handful among the 52 experts are concerned about these issues, it implies it will also surely become topic of debate beyond the working group. For some of these longer-term issues, it is already the case, with significant resources being devoted to research already. An absence of mention on that topic would discredit the whole expert group in the eyes of all those decision-makers turning to you for guidance in these debate. Seeing that you did not even mention the controversial issue, they might presume a telling lack of foresight on your behalf. Seeing that you satisfied yourself with mentioning these issues in a footnote, they might accuse you of a lack of courage. However, seeing that you explained in details the causes for concerns, the arguments and the assumptions underlying all sides of the controverse, and cautious, qualified guidelines, they will conclude you have shown intellectual honesty, humility, and a desire to inform the best you can for the benefit of society. In short, while the reasons for excluding these issues remain mostly logically invalid, I see at least these four reasons to include them : citizens' freedom to consider the issue themselves, the precautionary principle, the credibility of the EU as the global champion of ethics, and the population's need for guidance. Moreover, the risk from writing explanatory sections on these issues and the controverse is absent. If one of these issues turns out to be a silly idea to discuss, then that particular section will never be quoted, among the otherwise high-quality ethical guidelines your group provide. That's the ultimate, insignificant consequence. If however it turns out not to be a silly idea to discuss, then, a growing share of decision-makers will be made aware and inspired to act with caution thanks to you. That would be the significant legacy of your decision to elaborate on them.I wish you all the best for the thorny discussions ahead and, whether you agree or disagree with my point of view, I am grateful for all your efforts and best intentions.

Nadisha-Marie	Aliman		As a response to the invitation on sharing "thoughts on additional technical or non-technical methods that can be considered in order to address the requirements of Trustworthy AI", I would like to propose considering an alternative not yet mentioned Systems-Engineering oriented approach which especially incorporates an examination of the technical feasibility of given methods. As stated by Virginia Dignum, "artificial intelligence is no longer a computer science discipline, it is an interdisciplinary discipline". Along this line of thought, I believe that the AI Ethics and AI Safety fields can significantly profit from a holistic systems view as provided within a Systems Engineering oriented framework which inherently exhibits expedient interdisciplinary properties.In the following, I will briefly introduce this approach which might be able to jointly address a large number of requirements for Trustworthy AI formulated in this document including accountability, governance of AI autonomy, robustness, safety and transparency. First, for the purpose of accountability (but also safety, security and controllability), one might argue that it is in the interest of democratic societies to achieve an unambiguous and clearly formulated assignment of responsibilities with regard to the deployment of intelligent systems. Thereby, the systems should act in accordance with ethical and legal frameworks as specified by the legislative instance in order to facilitate the attribution of responsibilities by the judicial power. For the deployment of intelligent systems, this entails the necessity of a disentanglement of responsibilities for the "how" and for the "what" whereby the manufacturer of the systems are responsible for the "how" and the legislative for the "what". On a technical level, one therefore needs an approach able to practically realize that necessary disentanglement. Possible technical solutions would encompass a normative (rule-based) implementation or a consequential one (quantified in objective functions). However, for major technical reasons from which I will introduce the most important ones in the following, a consequential approach appears to represent the only feasible option. First, the attempt to try to formulate deontological rules for every situation an intelligent system might encounter in a complex real-world environment leads to a state-action space explosion (Werkhoven et al., 2018), which would not be the case in a consequential solution implementing a run-time adaptive utility maximizer exhibiting the properties of "self-awareness" (meant in a Systems Engineering sense and referring to: self-management, self-assessment and the ability to provide (symbolic) explanations) (Aliman and Kester, 2018). Utility maximizers represent a suitable solution in this context, because they incorporate the idea of having problem solving ability and ethical ability as orthogonal dimensions (as similarly formulated in the orthogonality thesis by Bostrom (2012)) which serves as a valuable feature if one wants to achieve the mentioned disentanglement of responsibilities for the "how" and the "what". Second, since law is formulated in natural language which is intrinsically ambiguous on multiple linguistic levels, either an intelligent system will have to extract meaning out of

this text material using fault-prone Natural Language Processing techniques or the developers might make use of ontologies encoding law which would however require them to first interpret law, which would in turn violate the idea of disentangling responsibilities. Third, legal frameworks often leave tradeoffs and dilemmas open which a normative approach cannot directly solve, a problem which a consequential system would not encounter. Fourth, an update of laws in the normative case will require every manufacturer to costly modify the built-in ethical framework, while the consequential solution would only require a centralised update of an ethical goal function* (Werkhoven et al., 2018) encoding the legal and ethical framework.In a nutshell, the presented analysis leads to the result that a technically feasible way to implement Trustworthy AI would be to disentangle the responsibilities for its deployment in such a way that the manufacturers are responsible for the second component of Trustworthy AI (technical robustness including safety and security) which they can implement using a consequential approach with "self-aware" utility maximizers**, while the legislative as representation of the whole society is responsible for the first component of "ethical purpose" by means of quantitatively specified and machine-readable ethical goal functions. By doing this, the following requirements for Trustworthy AI would be inherently addressed: accountability, governance of AI autonomy, robustness, safety and transparency. Moreover, the introduced approach offers the possibility to directly address the remaining requirements such as non-discrimination and respect for privacy by reflecting them in the mathematical formulation of the ethical goal functions.*https://www.tno.nl/nl/tno-insights/artikelen/we-dreigen-wereldwijd-achterop-te-raken/**A possibility to try to implement a utility maximizer equipped with a "self-awareness" functionality (self-management, self-assessment and the ability to provide explanations) could be to combine e.g. Deep Learning sensors at the subsymbolic level with a reasoner/planner on top of it at the symbolic level equipped with a causal model of the world and a self-model. The intelligent system performs actions (or a plan as sequence of actions) maximizing on utility given the ethical goal function.References:Aliman, Nadisha-Marie, and Leon Kester. "Hybrid Strategies Towards Safe “Self-Aware” Superintelligent Systems." International Conference on Artificial General Intelligence. Springer, Cham, 2018.Bostrom, Nick. "The superintelligent will: Motivation and instrumental rationality in advanced artificial agents." Minds and Machines 22.2 (2012): 71-85.Werkhoven, Peter, Leon Kester, and Mark Neerincx. "Telling autonomous systems what to do." Proceedings of the 36th European Conference on Cognitive Ergonomics. ACM, 2018.

Anonymous	Anonymous	Anonymous		Chapter II, 2.2 (page 22): "Non-technical methods to ensure ethical behavior of AI" As European Buddhists, we suggest that all developers and implementers within the EU of AI must be trained in ethics according to this document. Possibly, there might be even a certification process, which would be needed to obtain a "license" of some sorts. In simple words, the programmers and developers are akin to being „parents" of AI, and by what they "teach" is what the system "learns". Thus, if a developer should have few ethical standards, it will be difficult for him to implement those ethics into a learning system. Chapter II, 1.3 (page 15): We would like to underline the importance of thinking about clear criteria which support accessibility of the benefits of AI to all levels of society. Especially in health care, and possibly also education, the use of AI could lead to an even greater rift between wealthy and poor citizens, as many times new technology is also expensive at the beginning and hence only accessible to those who can afford it.		Thank you very much for this thorough and well-prepared document! We also appreciate very much the opportunity to give input.
JOSE	VARELA	UGT (Unión General de Trabajadores). Labor union of Spain.		For UGT, the only possible method to obtain a 'Trustworthy AI' is through the creation of public, independent and autonomous organizations for the inspection, control and audit of labour algorithms. It would be organizations with highly specialized and qualified personnel that could evaluate the decisions of the labour algorithms to audits, verifying the suitability of their operation. For example, to analyse the behaviour of any artificial intelligence platform, one must study the data set that has been used to train these algorithms. Thus, it would be verified if the data are biased or if the sample is sufficiently broad and significant to be truthful and plural, and therefore, completely objective. Finally, we do not want to miss the opportunity to remember that the current GDPR foresees that, in the case of the existence of automated decision-making, must provide "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject", always in order to verify if the operation of an algorithm is in accordance with law; therefore, we request no more than strict compliance with the European law in force.		The Spanish Union UGT wants to show its disagreement in the exclusion of any reference to the work factor, as an element to consider within the Draft AI Ethics Guidelines for Trustworthy AI, prepared by the High-Level Expert Group on Artificial Intelligence. Undoubtedly, the work factor is a key element for the development of free people in prosperous and advanced societies, such as those of the Member States of the Union. Labour represents a fundamental part of the European Welfare State, from an economic and fiscal point of view, as well as a key element for human dignity. In addition, evident labour rights derive from labour relations, as established in the Charter of Fundamental Rights of the European Union (Articles 5, 15 and 31, and by extension, 27 and 28), the Universal Declaration of Human Rights (Articles 23 y 24). and 24) and the ILO Conventions on Discrimination (Employment and Occupation, No. 111) and on Forced Labour (No. 29). It is indisputable that employment consolidates the individual's freedom, personal progress and full social inclusion. In fact, it is not acceptable to design ethics guidelines on the IA that do not contemplate the labor perspective, which affects the vast majority of our European citizens and is present in the daily life of our lives. It is not possible to imagine a future of employment governed by an AI that does not take into account basic aspects of humanity such as empathy, understanding, assertiveness, emotional intelligence, and of course, compassion and probity. We cannot allow a future of work piloted by autonomous decision algorithms, which can make decisions against basic principles of human dignity; that they take discriminatory, partial, irresponsible or even illegal decisions, from a penal and labour point of view.

							Consequently, we claim the inclusion, within the Project, of a specific chapter that studies the ethical guidelines of the AI in the work environment.
Marc	Rotenberg	Electronic Privacy Information Center (EPIC) (USA)	We write first to commend the EU AI High Level Expert Group for preparing the draft AI Ethics Guidelines and also for seeking public comment on the proposed draft. As we explain below, we strongly support many of the recommendations contained in the draft. We write also to bring your attention to the Universal Guidelines for AI, a similar policy framework, that could help clarify some of the issues in the draft AI Ethics Guidelines and address other issues that are not yet addressed in the draft Ethics Guidelines. The Universal Guidelines are intended to maximize the benefits of AI, to minimize the risk, and to ensure the protection of human rights. The Guidelines set forth twelve principles to guide the design, development, and deployment of AI: 1. Right to Transparency. All individuals have the right to know the basis of an AI decision that concerns them. This includes access to the factors, the logic, and techniques that produced the outcome.2. Right to Human Determination. All individuals have the right to a final determination made by a person. 3. Identification Obligation. The institution responsible for an AI system must be made known to the public.4. Fairness Obligation. Institutions must ensure that AI systems do not reflect unfair bias or make impermissible discriminatory decisions.5. Assessment and Accountability Obligations. An AI system should only be deployed after an adequate evaluation of its purpose and objectives, its benefits, as well as its risks. Institutions must be responsible for decisions made by an AI system.6. Accuracy, Reliability, and Validity Obligations. Institutions must ensure the accuracy, reliability, and validity of decisions. 7. Data Quality Obligation. Institutions must establish data provenance, and assure quality and relevance for the data input into algorithms.8. Public Safety Obligation. Institutions must assess the public safety risks that arise from the deployment of AI systems that direct or control physical devices, and implement safety controls.9. Cybersecurity Obligation. Institutions must secure AI systems against cybersecurity threats. 10. Prohibition on Secret Profiling. No institution shall establish or maintain a secret profiling system.11. Prohibition on Unitary Scoring. No national government shall establish or maintain a general-purpose score on its citizens or residents.12. Termination Obligation. An institution that has established an AI system has an affirmative obligation to terminate the system if human control of the system is no longer possible.There Universal Guidelines are available here: https://thepublicvoice.org/ai-universal-guidelines/There is also an explanatory memo and a list of references that accompany the Universal Guidelines (UGAI)	The approach to AI ethics based upon fundamental human rights commitments, which also underlie ethical principles, followed by operationalizing these values, is excellent. The human-centric approach to underscore basic rights of dignity, freedom, equality, and justice, and that a "human being enjoys a unique status of primacy in the civil, political, economic, and social fields" is fundamentally consistent with our UGAI principles of rights to human determination, identification, and termination. That is, not only do human beings have the right to this unique status, but when AI augments or replaces human decision making, humans have the right to ensure that there remains human control and accountability. We especially applaud and support section 3 (Fundamental rights of human beings) which draws from the EU Treaties and Charter, but fundamentally from the UDHR (which frames our UGAI guidelines too).This preamble on fundamental rights (Chapter 1) includes important principles regarding due process (justice), fairness (equality), and freedom from sovereign or government intrusion (profiling and unitary scoring) that we also address in four of our UGAI principles: • Section 3.2 (freedom from sovereign or govt intrusion) is consistent with two guidelines from the UGAI which address both the protection of human rights of freedom, as well as a minimization of scope in overbroad collection and use of data:• UGAI-10: Prohibition on Secret Profiling. No institution shall establish or maintain a secret profiling system.• UGAI-11: Prohibition on Unitary Scoring. No national government shall establish or maintain a general-purpose score on its citizens or residents.• Section 3.3 (justice and due process) is consistent with our second guideline regarding a right to human determination in algorithmic decision-making:• UGAI-2: Right to Human Determination. All individuals have the right to a final determination made by a person. • Section 3.4 (equality) is consistent with our fairness obligation guideline, to prevent bias or discriminatory outcomes from algorithmic processes.• UGAI-4: Fairness Obligation. Institutions must ensure that AI systems do not reflect unfair bias or make impermissible discriminatory decisions.	This chapter on the realization of trustworthy AI (Chapter 2) includes appropriate attention to elements of data integrity, quality, and the assessment of AI systems and processes, which we also address in three of our UGAI principles: • UGAI-5: Assessment and Accountability Obligations. An AI system should only be deployed after an adequate evaluation of its purpose and objectives, its benefits, as well as its risks. Institutions must be responsible for decisions made by an AI system.• UGAI-6: Accuracy, Reliability, and Validity Obligations. Institutions must ensure the accuracy, reliability, and validity of decisions. • UGAI-7: Data Quality Obligation. Institutions must establish data provenance, and assure quality and relevance for the data input into algorithms.These principles are essential to ensuring that AI systems and practices are technically robust and reliable.	The focus of Chapter 3 on assessment of trustworthy AI provides important guidance on evaluation of AI systems. What is not evident is a clear statement regarding the accountability for the outcomes and consequences of AI systems. Such accountability, including explainability of AI, is a critical obligation both legally and ethically, and should be clarified. Our UGAI includes language that would be helpful:• UGAI-5: Assessment and Accountability Obligations. An AI system should only be deployed after an adequate evaluation of its purpose and objectives, its benefits, as well as its risks. Institutions must be responsible for decisions made by an AI system.• Assessment determines whether an AI system should be established. Imperatively, such assessments must include a review of individual, societal, economic, political, and technological impacts, and a determination can be made that risks have been minimized and will be managed. Individual level risk assessments might include a privacy impact assessment; societal level risk assessments might involve public health or economic impact assessments. If an assessment reveals substantial risks, especially to public safety and cybersecurity, then the project should not move forward. Accountability for the outcomes and consequences of AI systems lies with the institutions.	We would like to indicate guidelines we think are important that are not yet addressed in this document. First is one of the three issues that have become universally accepted as important foundational principles in AI policy – transparency (along with fairness and accountability). The UGAI includes a statement on this principles that could be helpful:• UGAI-1: Right to Transparency. All individuals have the right to know the basis of an AI decision that concerns them. This includes access to the factors, the logic, and techniques that produced the outcome.• This principle of transparency, foundational in most modern privacy law, is grounded in the right of the individual to know the basis of an adverse determination. The obligation of transparency also serves the collective public, not only individuals who express specific harm. Assessment results should be made public to allow an opportunity for unknown biases to be made identified.In addition, issues of safety and security, especially in areas of AI development such as transportation and national defense industries, must be addressed as well. The UGAI includes two specific guidelines that can be helpful here:• Public Safety Obligation. Institutions must assess the public safety risks that arise from the deployment of AI systems that direct or control physical devices, and implement safety controls.• Safety and security are fundamental concerns of autonomous systems – including autonomous vehicles, weapons, and device control – and risk minimization is a core element of design. Less certain, however, is how to determine and set standards for levels of autonomy across broad applications, and understanding levels of autonomy (and the correlate level of human control) is an interdisciplinary research challenge. The UGAI underscores the obligation of institutions to assess public safety risks that arise from the deployment of AI systems, and implement safety controls. • Cybersecurity Obligation. Institutions must secure AI systems against cybersecurity threats. • Institutions must secure AI systems against cybersecurity threats, particularly in the case of systems that act autonomously, such as autonomous weapons and vehicles, but also in the case of technologies that interface with or are embedded within humans. Even well-designed systems are vulnerable to hostile actors, and minimization and active management of such risks is a critical obligation. Finally, coupled with safety and security is a final principle regarding the assurance that human control of AI systems. We address this principle in our twelfth and final guideline:• UGAI-12: Termination: In addition, the final principle in the UGAI

			available here: https://thepublicvoice.org/ai-universal-guidelines/memo/In the sections below, we provide our comments on the various chapters of the draft AI report and also suggest how it may be possible to incorporate some of the text from the Universal Guidelines (UGAI).				states that institutions that have established an AI system have an obligation to terminate the system if human control of the system is no longer possible. This ultimate statement of accountability addresses not only autonomous systems, but also decision-making or decision-support systems that have been assessed. It is essential to ensure the safety and security of people, and research strategies need to address the development of assessment tools to determine loss of autonomy, alongside understanding the underlying question of what level of autonomy is appropriate for specific applications and contexts.
Giuseppe	Attardi	Università di Pisa	The document confuses AI with AI systems, sometimes using one (92 occurrences) or the other (123 times) interchangeably. But AI is a discipline (as mentioned in the glossary) whose purpose is to study models and techniques that can be used to build intelligent systems. It makes no sense to say that a discipline is trustworthy. The document should use "AI systems" throughout.	I disagree with the use of the term "AI Ethics". AI Ethics would mean a branch of ethics based on AI. One should talk instead of the "Ethics of AI" (https://en.wikipedia.org/wiki/Ethics_of_artificial_intelligence), clarifying that these guidelines consider "the moral behavior of humans as they design, construct, use and treat artificially intelligent beings", since we are not ready yet to discuss "the moral behavior of artificial moral agents". "Human-Centric AI" is poor wording. "Human-centric" is applicable to a design or to an approach to design, it is not a property of a discipline, like AI is. Besides, not all AI systems need a human-centric design since not all AI applications are directed towards the citizens.	Most of the requirements of Section 1 would apply to any information system, in particular accountability, non-discrimination, respect for privacy, robustness, safety and transparency. The document should focus on those that are additional or specific to AI systems. Otherwise AI systems will be encumbered with extra burden: for example automated lending systems which are statistically based are also subject to bias in data. Common software licences today provide no liability and take no responsibility for consequential damages. Nobody would produce AI systems if they would be liable of consequential damages. It should be always humans who are ultimately responsible and accountable, not AI systems, at least until we find a way to punish them. Section 2. Data governance. The measures for ensuring properly training and validation of ML systems are standard practice in any text book, they are a necessity, not an extra requirement for trustworthy AI systems. 3. Design for all Saying that the design should "allow all citizens ...", stresses the fact that the guidelines apply ONLY to systems with which user interact directly. For example, an AI system to optimize energy consumption in a datacenter need not fulfill this requirement. Section 5. Non-discrimination When saying that "Those in control of algorithms may intentionally try to achieve unfair outcomes" illustrates a confusion in the document about who is responsible. Responsibility is personal: therefore if someone designs and sells a malevolous software, he is liable, irrespective of whether the software uses AI or not. Section 8. Robustness "Resilience to attacks. AI systems, like all software systems", here the experts recognize that AI system do not differ from other software (why not hardware though?). They however immediately insist in their bias, saying that "if an AI system is attacked, the data as well as the system behavior can be changed". Once again, nothing special about AI, any database systems suffers the same risks.	The Guidelines are too biased towards Statistical Machine Learning, which are not all of AI. Most of the concerns are related to issues due to training, datasets, explainability, etc, which are typical of this approach but not of others. The guidelines, as formulated, should apply ONLY to "AI systems that relate ... to humans", as stated on pag. 24. This should be stated clearly at the beginning. AI is a discipline and therefore it makes no sense to attribute certain properties to a discipline, like "human-centric" or "trustworthy". No one would talk about "human-centric physics" or "trustworthy chemistry". "human centric" might apply to an approach to design of systems. A trustworthy discipline might be one whose methods are scientifically based and verifiable, but that is a truism for AI.

			Fall back plan. Once more, how many times we would have liked that an automated systems that fail to work as expected or are incapable of dealing with a certain situation could have a fall back on a human operator (press 9 to talk to an operator)? I have dozed of examples in my personal experience. Nothing special about AI systems. 10. Transparency Explainaability is the sort of "motherhood and apple pie" of discussions on AI. I agree completely with Geoff Hinton, who argues that we do not even request explainability for humans: "One place where I do have technical expertise that's relevant is [whether] regulators should insist that you can explain how your AI system works. I think that would be a complete disaster. People can't explain how they work, for most of the things they do. When you hire somebody, the decision is based on all sorts of things you can quantify, and then all sorts of gut feelings. People have no idea how they do that. If you ask them to explain their decision, you are forcing them to make up a story." On pag. 19, the experts suggest to "formulate rules which constrain the behavior of an intelligent agent" and that are run "in a separate process". The experts should know that formulating rules is as difficult and error prone to building an AI system. There is no guarantee that such rules would be capable to cover all situations and that they would work in conjunction with a system whose critical parts are based on AI. For example, a rule that says that a car should not run over a person, in order to be verified, must rely on a vision system that recognizes humans. On pag. 20. Testing and validation The document suggests that "action commands "be compared to the previous defined policies to ensure that they are not violated". As mentioned earlier, this is technically impossible, since program verification is undecidable. Moreover, if one had a system capable of verifying the correctness, it could be reversed in order to produce a solution. Therefore AI systems can be tested and validated by no better means than any other software system.
Miguel	Luengo-Oroz		Solidarity is one of the fundamental values of Europe and is at the heart of European construction. Solidarity can be understood as sharing both the advantages, i.e. prosperity, and the burdens equally and justly among members. Solidarity is the Title IV in the Charter of Fundamental Rights or the European Union and is core of other European mechanisms such as the European Solidarity Corps or Solidarity Fund. It is possibly the concept that makes Europe unique vis-à-vis other regions of the world and gives European citizens unique standards of social protection and wellbeing. In the current draft, the word "solidarity" is just mention in the title "3.3 Equality, non discrimination and solidarity", however that paragraph just explains about equality.

					There is no attempt in the draft to elaborate how European solidarity will shape the European angle towards the future of artificial intelligence. In fact such concept in the core of the AI developed in Europe could have multiple implications. As Yuval Noah Harari explains, the main struggle in the 21st century will be about irrelevance. AI will augment productivity at cost of human jobs and as he quotes "just as mass industrialization created the working class, the AI revolution will create a new unworking class". In the current state of the game, whoever has compiled enough data to make an AI model will be able to replicate an action almost for free. This will provide an advantage without precedents to the first one that gets it. Something that big companies and some governments know (and are pursuing in the middle of the confusion). It has been described as a new cold war. Europe needs to think upfront how we will redistribute the productivity gains due to AI (automation in its basic level). Those whose data can be used to train the models could receive some reward in exchange, or automated tasks should pay taxes. Solidarity should be implemented in form of mechanisms to exacerbate the creativity and relevance of humans, and tackle the key challenge of redistributing some of the augmentation of productivity provided by AI, so we attempt to equalize a world each day more unequal. At a different level, another practical implementation of the solidarity concept could be to promote open source developments. Were the AI HLEG to consider to include solidarity as one of the core Ethical Principles in the context of AI and correlating Values or to elaborate this concept in other part of the document, I'll be happy to discuss and contribute further.		
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Anonymous	Anonymous	Anonymous		• While the concept of ethical purpose is well documented (and mentioned 25 times in the entire document), the concept of technical robustness (only mentioned 3 times) should also be clarified and its relevance in this context should be explained. • The Principle of Autonomy: "Preserve Human Agency" (p. 9): about the right to opt out: AI will be everywhere and it will be impossible to opt out of all AI systems. Only a small part of AI systems will be relevant enough to insist on having a right to opt out, e.g. if there is a strong impact on human dignity (e.g. euthanasia). The cost for society (government and business) will be too high if many people would opt out of all AI systems and it would be economically impossible to provide human based alternatives for everything. Only for some specific cases should an opt out option be given. On the other hand, we should insist on having the possibility to appeal to an automated decision, even if the appeal is handled by another system. As decisions by AI systems will have long term impact on the life of everyone, we should ensure these decisions are fair, rather than offering to opt out. • The Principle of Justice: "Be Fair" (p. 10). I think there's more to justice and fairness in	• 1. Requirements of Trustworthy AI (p. 14): It isn't clear how the requirements are derived from the rights, principles and values from Chapter I. At least one example should be given to understand the logic. • 1. Accountability (p. 14). Clarifying ex ante and ex post how accountability / moral responsibility (see p.8 of the EGE document) deserves more attention in the guidelines. Given the many hands problem, the role of humans (e.g. in the loop), dual use and potential abuse, it is rather complex and confusing, though crucial to understand. In each of the 4 use cases, this could be elaborated. • 6. Respect for (& Enhancement of) Human Autonomy (p. 16): Manipulative nudging should not be used as a term. Nudging is per definition OK (a little paternalistic push without limiting the freedom), manipulation is per definition questionable. AI should actively be used to nudge people. Some cases of manipulation are OK (e.g. marketing will manipulate people to buy products and services), other cases aren't OK (e.g. in the context of elections).	• The items in the Assessment List are broadly derived from the requirements in chapter II. A more systematic mapping of all aspects of the requirements would make the list more comprehensive. • p. 24: "It will include specific metrics, and for each metric key questions and actions to assure Trustworthy AI will be identified. These metrics are subsequently used to conduct". I see many questions, but no metrics, nor actions. At least one example with metrics, questions and actions should be given. • 6. Respect for Privacy (p. 25): The proposed questions are focused on compliance with regulation/legislation. Are there no ethical questions that go beyond?	• Is the mandate as defined in the concept note of the EC fully covered? "Propose to the Commission AI Ethics Guidelines, covering issues such as fairness, safety, transparency, the future of work, democracy and more broadly the impact on the application of the Charter of Fundamental Rights, including privacy and personal data protection, dignity, consumer protection and non-discrimination". Concrete deliverable 1: draft AI Ethics Guidelines • We could copy the approach of GDPR mutatis mutandis, with an AI impact assessment (already exists in Holland https://ecp.nl/jaarcongres/artificial-intelligence-impact-assessment/; cf. DPIA), an Ethical AI Officer (cf. DPO), X by design (cf. privacy by design), ethical policy (cf. privacy policy), National Ethical AI Office (cf. DPA), etc. • For situations in which there's the risk of people dying as a consequence of a decision by AI (e.g. self driving car, some medical decisions), we could insist on having a device similar to the black box in the aviation industry (flight recorder) that helps understand ex post what has happened. • Similar to the aviation industry, we should have an institution (e.g. the National or

				this context. E.g. with a Rawlsian approach, rather than 'evenly distributed' positives and negatives. • 5.2 Covert AI systems (p. 11): I believe that AI will be embedded in all parts of our environment and that we will expect AI to be everywhere (e.g. who thinks Siri is a human being?). So I don't see the need to avoid covert AI, on the contrary, good AI is invisible. There is one exception: AI should not be used for impersonation, e.g. an AI version of a political candidate who could call thousands of people to ask them for their vote, should not pretend to be that candidate but at the start identify itself as an AI aid, even if the AI mimics the way of speaking of the politician.		European Ethical AI Office) that investigates all lethal cases with as main objective to learn from what has happened to avoid similar situations in the future. It should make its recommendations public to all stakeholders involved with AI. • What about recommending a methodology like Value Sensitive Design? • What about proposing an hippocratic oath for ethical AI? If not at European level, maybe at national, industry or organisational level? • The document refers 7 times to secure AI, together with cutting edge and ethical. It isn't clear why it is so important compared to all other aspects that are mentioned, why it is mentioned on top of being ethical suggesting there is insecure ethical AI, and there isn't a clear definition of what secure means in this context. • Applicable regulation is mentioned 7 times but there isn't a clear explanation as why it is at the same level as fundamental rights and core principles and values and how it relates to the content of this document. • I believe the terminology should be more consistent. • Inconsistent use of terminology related to the lifecycle, following terms are used in various combinations: development, deployment, use, design, implementation, application, realisation, research, regulation, execution, evaluation, analysis, justification, data gathering, testing, maintain, training, usage, identifying requirements, evaluating solutions, ensuring improved outcomes • Inconsistent use of terminology related to the stakeholders: users, investors, innovators, managers, developers, employees, deployers, designers, consumers, implementers, general public, stakeholders, social partners, actors, governments, companies, organisations, researchers, public services, institutions, individuals, other entities • Inconsistent use of terminology related to the beneficiaries: the common good, individuals, well being, citizens, human beings, communities, groups of people, minority groups, vulnerable groups • Inconsistent use of terminology related to systems in scope: AI, AI-based systems, intelligent systems, learning systems, neural nets, neural networks, algorithmic system • There are several concepts mentioned related to explicability. It isn't always clear what is meant by each of these and whether they are used in a consistent manner: explicability, explainability, intelligibility, transparency, accountability, explanation, auditable, auditability, comprehensible
John	Zerilli	University of Otago			Item 4 of the Assessment List in this chapter includes the following two questions: "In what ways might the AI system be regarded as autonomous in the sense that it does not rely on human oversight or control? What measures have been taken to ensure that an AI system always makes decisions that are under the overall responsibility of human beings?" Both these questions appear to have an "ultimate" notion of control in mind, the ability of a human controller to redirect or terminate the operations of a machine. This	Overall, this is a well-crafted and humane document that sets out the relevant issues in a clear and thoroughgoing way.

						is fine. But another notion of control might be called "proximate", the sort of control that means a human, in addition to having ultimate/overall control, has control over the execution of discrete steps in a machine's operations. The phenomenon of automation bias, and automation complacency, however, makes it undesirable for a human always to have this sort of micro-level control. If a system performs more reliably than a human, research indicates that human interference can actually make the system worse. So while we want humans ultimately in control of a machine, in the sense that we should be able to terminate its operations if we so choose, this does not entail that we should have control over every, or most, or even any, steps in its actual execution of a task. I think the report should be a little clearer on the sort of human control it is recommending. If the report's writers wish to advocate what I have called proximate control, over and above ultimate control, then the report will need to have something to say about how it can avoid the well-researched problems of automation bias and complacency. We have a clear summary paper on this issue, if you would like to know more.	
Matt	Allison	Vodafone	Vodafone welcomes the opportunity to provide comments to the European Commission's High Level Expert Group AI (AI HLEG) on its Draft Ethics Guidelines for Trustworthy AI. In preparing this work, the Commission has identified three pillars to support and promote the development of ethical AI in the EU: boosting investment, preparing for socio-economic changes and ensuring an appropriate ethical and legal framework to strengthen European values. In our view these pillars correctly address the most urgent policy challenges emerging from the global race to adopt AI technologies. For European policy makers, these challenges are particularly acute, and the way in which we respond will determine the extent to which Europe leads or languishes in the race to become a global AI powerhouse. Our strongest rivals, the US and China, benefit from economies of scale stemming from large domestic markets under a unified regulatory framework. In addition, the balance in both of these markets between unchecked innovation and upholding individual rights is skewed towards the former, placing less restrictions on what AI developers can do within the regulatory framework. Lastly, and perhaps most importantly from the perspective of boosting AI innovation, in both the US and China funding for AI research and development is being made available on an enormous scale. While in the US much of this funding derives from private equity and Venture Capitalist activity, in China the centrally planned economic system has guaranteed significant sums to be made available for AI development under the 'Made in China 2025' economic plan. To compete on a global scale, the EU must be able to leverage its own resources and those of member states national exchequers to ensure that equivalent capital is being invested in indigenous EU AI technology. Correspondingly EU policy makers should	General comments To the best of our knowledge there is no widely accepted international guidance available to determine what is ethical and what is not. Human Rights Conventions and national laws based on these conventions are at present the only commonly agreed articulation of what is ethical. Therefore, ethics should be based on them. Vodafone has conducted an internal mapping of Human Rights Conventions and data processing practices for our internal compliance documentation (not specific to AI). This has been done in context of trying to understand what are the "rights and freedoms of Individuals" that may be impacted by various data processing. We commend this approach (which incidentally has the support of the EUDPS) to the AI HLEG and suggest that this could become a general requirement for companies deploying AI technology.Vodafone supports the adoption of a self-regulatory approach to AI development in the EU grounded upon strong fundamental rights. The introduction of the General Data Protection Regulation demonstrated the ability of EU policy makers to intervene to uphold EU citizen's fundamental rights beyond the EU's borders. The extra territorial effect of the GDPR, combined with the sheer size of the addressable market for online goods and services within the EU has led to a rapid adoption of GDPR like data protection standards across the globe. As such the EU has been able to leverage its scale and soft power to shape the global market for data driven products and services. Ethical AI presents a similar opportunity for the EU to create the global gold standard for secure and trustworthy AI. Vodafone endorses the overarching conceptualisation of the relationship between fundamental rights, principles and values set out at the beginning of this chapter. The cascading relationship from fundamental rights (legal, immutable) down to ethical principles	General comments Mapping the list of requirements for trustworthy AI against Vodafone's internal policy and practices we see a high degree of commonality with the Vodafone approach (our own approach centres on Transparency & Accountability, Ethics & Fairness, Security and Privacy, Humanity and Diversity). The AI HLEG list is somewhat repetitive: it should be made clear in this document if these requirements are ranked in order of importance.An alternative list could perhaps be structured slightly differently and could include some other items: 1. Procedural requirements:o Accountability and Human oversight;o Data Governance;o Impact assessment and Ethics/Fairness by Design (both for new AI technologies and for each new use case of existing technologies, including regular re-evaluation of existing technologies and use-cases)o "Conservative approach" - i.e. if something is tried out for the first time, one should first try it out on a very limited audience, subject to highest level of human oversight. 2. Material requirements (i.e. "criteria" for impact assessments):o Respect of international human rights laws and standards o Legitimacy and compliance with lawso Transparencyo Human agencyo SecurityWe may also want to consider the extent to which some kind of public oversight or governance mechanism is needed (regulatory, judicial or perhaps a parliamentary committee that publishes guidelines). Detailed comments AI HLEG text (Data Governance pp 14): The datasets gathered inevitably contain biases, and one has to be able to prune these away before engaging in training. This may also be done in the training itself by requiring a symmetric behaviour over known issues in the training setVodafone comments: 'Data pruning' is an interesting concept, and one that should be further elaborated on by the AI HLEG. Is there a risk of over deletion/overzealous pruning? We would also appreciate	General commentsVodafone is supportive of the objective of this chapter of the ethics guidelines, to provide a practical checklist of questions to consider in order to ensure the development of trustworthy/human centric AI from an early stage of the development cycle. However, our concern with this section is that while the questions here are appropriate considerations, they lack the technical detail and specificity which would make them a useful or practical tool for AI product development or engineering staff, particularly within a small organization. In general, there is too much repetition of the early sections of the ethics guidelines and not enough effort to provide a clear structure and benchmarking/self-assessment tool for AI developers which is easily understandable in a variety of different languages and AI use cases. Detailed commentsAI HLEG text (Data Governance pp 24): What data governance regulation and legislation are applicable to the AI system? Vodafone comments: We would ask the AI HLEG to consider carefully the linkages which exist between the debate around AI ethics and the ongoing debate around establishing a level regulatory playing field for equivalent services, particularly in the context of the ePrivacy Regulation. Vodafone advocates for a proportional regulatory framework around data to enable AI innovation, with a level regulatory playing field between operators. Vodafone also supports the development of globally harmonised data protection standards, based on free and open data flows to prevent the imposition of unjustified data localisation requirements which would certainly hamper the development of AI. We suggest the AI HLEG could include specific language here on the need for more harmonised cross border data transfers standards i.e. by referring to existing EU standards such as GDPR and the Free Flow of Data regulation. However, Vodafone would strongly reject the introduction of a	General CommentsAt Vodafone, we are using AI to help to improve our products and services and to run our business as effectively as possible: AI chat bots increase the speed with which we can respond to routine customer enquiries; our 'big data' team uses AI to analyse large, anonymous data sets from customers (who have given us permission) so we develop new and better products and services; and we are deploying AI technology in our mobile networks to identify where capacity is needed so our customers can make calls and access the internet without interruption.As AI grows in usage and impact, we have a responsibility to consider how our use of this technology impacts our customers, our employees, and wider society. We believe it is critical to ensure that the AI algorithms we use are designed to respect both the privacy and security of the data they analyse. We also want to ensure that the insights we derive from big data are fair and not subject to any unintended bias.

attempt to ensure a harmonized regulatory environment across the EU with regards to AI technology to avoid costly fragmentation. We do not envisage that the EU's rights based framework should be a disadvantage to AI innovation, or hamper the EU's ability to compete at the global level. Rather we agree with the European Commission that a strong ethical framework, grounded in fundamental human rights underwritten by the EU's founding treaties, could be leveraged for competitive advantage, by making the EU home to trustworthy AI solutions. Only AI which is trustworthy and robust has the potential to achieve mass market adoption: thus by positioning the EU as the standard bearer of Trustworthy AI from the outset, we are confident that EU technology providers can become the global exporting powerhouses of the future. We need only observe the global reaction to the introduction of the General Data Protection Regulation (GDPR) to see how the exercise of 'soft power' based on a strong rules based framework can indeed help to shape global markets to the EU's competitive advantage. We see no reason why AI ethics should be any different, and as a leading European technology company, Vodafone is strongly committed to working with the European Commission and multistakeholder groups to deliver on this ambition. Our comments on the draft ethics guidelines are intended primarily to ensure consistency with existing regulatory frameworks, and to advise where measures proposed by the AI HLEG are either disproportionate or technically not feasible. In all cases we have suggested alternative wording which should bring the ethics guidelines into line with industry best practice, and ensure a clearer link with AI developments which are currently underway across the telecommunications ecosystem. Detailed comments AI HLEG text (Executive Summary, pp III): Trustworthy AI has two components: (1) it should respect fundamental rights, applicable regulation and core principles and values, ensuring an "ethical purpose" and (2) it should be technically robust and reliable since, even with good intentions, a lack of technological mastery can cause unintentional harm.Vodafone comments: Vodafone considers the concept of 'technological mastery' to be too vague. Suggested alternative: "(2) it should be technically robust and reliable, making use of international cybersecurity standards and best practices to eliminate unintentional vulnerabilities". AI HLEG text (Executive Summary, pp III): In contrast to other documents dealing with ethical AI, the Guidelines hence do not aim to provide yet another list of core values and principles for AI, but rather offer guidance on the concrete implementation and operationalisation thereof into AI systems. Such guidance is provided in three layers of abstraction, from most abstract in Chapter I (fundamental rights, principles and values), to most concrete in Chapter III (assessment list).Vodafone comments: The document describes itself as an operational/practical set of guidelines for developers of AI systems, however it suffers from many of the same problems of other industry standard AI policy/ethical guidelines: i.e. it lacks technical specificity and aspires towards high level principles rather than granular commitments against which (abstract, high level norms to uphold human centric/trustworthy AI) and finally to values (granular, measureable guidelines for a business to operationalize those rights and principles) is a helpful one.In operationalising these guidelines, providers of AI technology should be empowered to contextualise and make adjustments to suit the various different technology use cases they deploy. Vodafone defines Artificial Intelligence as 'the application of advanced analytical techniques (ML, DL and NLP) combined with automation and related feedback loops to solve problems and seize opportunities in new ways". Our use of AI falls into two broad categories: technology focused AI and commercially focused AI. For the former, AI is being used to assist with fault detection, predictive maintenance, networking planning and optimisation, all of which combines to ensure that we are able to make more efficient use of our physical assets. With regards to the latter, AI is also being used in a commercial setting, through the deployment of Virtual Assistants (more information on our chatbot Tobi can be found here), pricing and promotions, predictive care, smart retail and more. The AI HLEG should establish at the outset that a 'one size fits all' approach is not suitable for AI applications, and that a determination should be made on a case by case basis how ethics guidelines should apply to different AI use cases. Detailed comments AI HLEG text (Respect for Human Dignity, pp 7): To specify the development or application of AI in line with human dignity, one can further articulate that AI systems are developed in a manner which serves and protects humans' physical and moral integrity, personal and cultural sense of identity as well as the satisfaction of their essential needs.Vodafone comments: Advocating AI which serves and protects human's moral integrity could perhaps be too onerous for many AI use cases. The concept that neutral technology should be able to 'protect' an individual appears potentially problematic, particularly in scenarios where the AI being deployed has no direct interface with the end user (for example AI technology deployed within an ECS network for the purpose of fault detection and remedy). Suggested rewording: 'one can further articulate that AI systems are developed with full respect to human's physical and moral integrity'. We do acknowledge however that in some circumstances AI may be used to replace human interaction where it may have to undertake a human judgement task. In such cases it is necessary for the agent to preserve and follow established ethical norms for preserving human dignity in any form.AI HLEG text (Citizens Rights, pp 7): At the same time, citizens should enjoy a right to be informed of any automated treatment of their data by government bodies, and systematically be offered to expressly opt out.Vodafone comments: In our view this requirement in the ethics guidelines goes too far. Automated decisions involving personal data are already subject to GDPR requirements. AI HLEG text (Ethical Principles in the Context of AI and Correlating Values, pp 8): in particular situations, tensions may arise between the principles when considered from the point of view of an individual compared with the point of view of society, and vice versa. There is no set way to deal with such trade-clarification from the AI HLEG on the extent to which this practice differs from "data minimisation" - an established concept under GDPR. Is it perceived as an ongoing process, to check and re-check the data volume/breadth over time? The principle of Human Rights Impact Assessment is critical here. By asking the right set of questions, the potential biases will be more likely to surface. Importantly, while bias is typically driven by data quality, mechanisms can be set up to minimize bias also through the algorithms themselves.AI HLEG text (Data Governance pp 14): Feeding malicious data into the system may change the behaviour of the AI solutions. This is especially important for self-learning systems. It is therefore advisable to always keep a record of the data that is fed to the AI systems, to the extent possible within technical and regulatory constraints (for e.g. GDPR requirements with regard to data erasure). Vodafone comments: Vodafone supports a strong requirement on digital platform providers to ensure appropriate control over the input data being served to AI, to ensure that it is neither malicious, nor in breach of hate speech laws or societal norms. We recognize however that it is very difficult and impractical to monitor real-time data and ask customers for example to use a certain type of language only. Instead of training customers, a reinforced learning system should be trained to either ignore or not use certain type of data sets, backed up by strong human oversight. AI HLEG text (Non-discrimination pp 16): It is important to acknowledge that AI technology can be employed to identify this inherent bias, and hence to support awareness training on our own inherent bias. Accordingly, it can also assist us in making less biased decisions.Vodafone comments: Vodafone strongly supports the reference to the power of AI to help society identify and eradicate inherent biases and to assist humans in making less biased decisions. The AI HLEG could commit to a more detailed examination of the positive examples of AI being used to tackle inherent bias. Establishing respect for Human Rights as a fundamental parameter of the underlying technology should help produce this result.AI HLEG text Respect for (and enhancement of) Human Dignity: AI products and services, possibly through "extreme" personalisation approaches, may steer individual choice by potentially manipulative "nudging". At the same time, people are increasingly willing and expected to delegate decisions and actions to machines (e.g. recommender systems, search engines, navigation systems, virtual coaches and personal assistants).Vodafone comments: Personalised AI solutions have the potential to vastly improve the consumer experience, saving them money and offering them timely and relevant deals. We are aware however that consumer IoT devices, linked to ubiquitous digital platforms (home assistants) have the potential to influence user behaviour and diminish human autonomy and dignity. Providers of consumer IoT products and online intermediation services should not be able to leverage this technology to unfairly promote or benefit their products and services. In addition, we suggest that the AI HLEG focus on the following key principles: i) transparency as a tool to ensure consumers horizontal "data governance law". Different data is subject to different legal protections arising from a number of different sources. Data governance are the internal practices that ensure those rules are complied with and not something which requires external regulation. Market mechanisms should govern access to data held by private entities, not regulation. From an economic perspective, data is becoming a valuable asset that underpins innovative data-driven business models. Mandatory requirements for data to be made available to public authorities would be a disincentive to invest in technology that would generate data in the future, and would therefore act as a brake on innovation. Operators investing in tools and technology to collect and analyse data should be able to develop a commercial model to for the reuse and aggregation of this data. Sharing should take place on the basis that it is legally valid, socially acceptable and economically viable.AI HLEG text (Governing AI Autonomy pp.24): Is a process foreseen to allow human control, if needed, in each stage? Is a "stop button" foreseen in case of self-learning AI approaches? Within the organisation who is responsible for verifying that AI systems can and will be used in a manner in which they are properly governed and under the ultimate responsibility of human beings?Vodafone comments: This is a critical question. In our view, it is not necessary to guarantee human control at all levels of AI (e.g. where this is deployed deep in the network for fault detection) Human control may in some cases only be necessary in setting the outcomes, whereas in other cases, where there is a significant impact on individuals, human control is essential. The "conservative approach" and "human oversight" principles should certainly help with this objective. Human control and a stop button failsafe may not be necessary precautions for all types of self-learning AI approaches. Indeed, if we mandate these types of control, even for AI which is deployed deep in our networks and has no human interaction or customer facing element, we could deprive AI of its greatest potential: to solve problems (like cancer treatment, reversing global warming etc.) which humans have not been able to. We request that the AI HLEG give more detailed thought to some of these subsidiary questions before including human control/stop button as a requirement in this section of the guidelines. A contextual understanding of AI, where different use cases are permitted differing levels of human control appears to us to be the optimum outcome.

companies and individuals can be audited. This may ultimately be the best approach, but we should be careful about the AI HLEG selling this document as something it is not. AI HLEG text (Executive Summary, pp III): In the final version of these Guidelines, a mechanism will be put forward to allow stakeholders to voluntarily endorse them.Vodafone comments: Vodafone would welcome additional detail from the Commission on what the endorsement/certification mechanism will entail in the final draft guidelines. In particular, we would welcome clarity on how this mechanism will dovetail with existing and forthcoming codes and quality schemes at the national, EU and global level. The AI HLEG would be advised to avoid duplication of existing schemes where possible to avoid confusion for consumers of AI solutions and unnecessary complexity for AI developers. AI HLEG text (Executive Summary, pp iv): Moreover, the Guidelines should be seen as a living document that needs to be regularly updated over time to ensure continuous relevance as the technology and our knowledge thereof, evolves. This document should therefore be a starting point for the discussion on "Trustworthy AI made in Europe".Vodafone comments: Vodafone strongly supports this approach. We believe that to remain relevant this should be a living document, updated over time to reflect new AI insights and market developments. A practical model for this could be to split the guidance into rights/principles and practical guidance levels. The former should remain stable while the latter should be revised in light of technological and market developments at appropriate intervals (every twelve to eighteen months). AI HLEG text (Executive Summary, pp iv): Finally, beyond Europe, these Guidelines also aim to foster reflection and discussion on an ethical framework for AI at global level.Vodafone comments: Vodafone supports international harmonisation with regards to the rules governing ICT and Internet policy wherever possible. We ask whether the AI HLEG may go further than the above statement, in pushing not only for 'reflection and discussion' at the global level, but inviting international partners to adopt the EU model for human-centric/trustworthy AI, either through bilateral partnerships (the recent France-Canada AI declaration) or through multilateral norm and standard setting bodes (ETSI, ISO, GSMA, ENISA).

offs.Vodafone comments: The distinction between principles considered from the point of view of the individual, compared with that of society is a very interesting and complex one, and warrants greater attention than is granted in this document. Service providers like ours will need to think carefully about how they should act when there is a clear ethical tension between the needs of an individual user and society at large, and should have in place at the very least an ethical code or set of instructions to guide their behavior in these circumstances. It Is our experience that where such tensions exist, it should not just be left to private operators to make a determination on where the correct balance lies between fundamental rights. Such determinations should only be made by actors with a clear public mandate to decide where the appropriate balance should lie; usually judicial authorities or elected officials. Where private operators can play a role is through the introduction of technical measures which reduce the need for difficult tradeoffs: in the context of the ePrivacy Regulation pseudonymisation has been identified as a technical measure which can be deployed to ensure electronic communications data can be used without impinging on fundamental rights. We would encourage the AI HLEG to undertake a detailed study of whether technical measures exist which could reduce tensions between the rights or individuals and society at large as outlined above. AI HLEG text (The Principle of Beneficence: "Do Good" pp 8): AI systems should be designed and developed to improve individual and collective wellbeingVodafone comments: We are fundamentally concerned about the concept of AI systems being employed only for 'good'. 'Individual and collective wellbeing" is not well-defined concept in European Charter of Fundamental Rights, which is the relevant governing legal document here. On a practical level, we believe this could go beyond the remit of private actors, whose main responsibility is towards their shareholders and their customers, as governed by contract. As set out above, use of AI may simply be part of the technical evolution of communications networks and have no "good" or "bad" consequences, in the ethical sense. Vodafone asks the AI HLEG to give more careful thought to these difficult practical and philosophical questions before including a principle of AI beneficence in the final ethics guidelines. AI HLEG text (The Principle of non-maleficence "do no harm" pp 9): By design, AI systems should protect the dignity, integrity, liberty, privacy, safety, and security of human beings in society and at work.Vodafone comments: Concerns regarding AI duty to 'protect' noted herewith (Respect for Human Dignity, pp 7). We refer the AI HLEG to the UN Guiding Principles Reporting Framework which enshrines a duty for states to protect human rights and corporate entities to respect human rights. Suggested amendment: "AI systems should respect the dignity, integrity, liberty, privacy safety and security of human beings"AI HLEG text (The Principle of non-maleficence "do no harm" pp 9): AI systems should not be designed in a way that enhances existing harms or creates new harms for individuals.Vodafone comments: This sentence appears to be missing a vital qualifying concept of intent. Suggested

are empowered to make the right choices and ii) use of AI as a tool itself to empower consumers, a concept which has been further elaborated by ARCEP AI HLEG text (Technical and Non-Technical Methods to achieve Trustworthy AI): An evaluation of the requirements and the methods employed to implement these, as well as reporting and justifying changes to the processes, should occur on an on-going basis. In fact, given that AI systems are continuously evolving and acting in a dynamic environment, achieving Trustworthy AI is a continuous process.Vodafone comments: Vodafone strongly supports the focus on continuous and ongoing assessment of trustworthy AI systems. Trustworthiness is not a static concept, or something which can be guaranteed from one product innovation cycle to the next. AI HLEG text (Ethics and rule-of-law by design pp 19): Whenever an AI system has a significant impact on people's lives, laypersons should be able to understand the causality of the algorithmic decision-making process and how it is implemented by organisations that deploy the AI system.Vodafone comments: We ask the AI HLEG to clarify on whom should the obligation fall to ensure AI systems are intelligible to laypersons? Service providers should be obliged to be transparent and up front about how AI is being deployed across their networks however we would argue that responsibility must sit with governments and public authorities to ensure a higher level of education in relation to how AI works (a basic standard of AI literacy could be a feature of a forward looking educational curriculum). This can only be built up in time just as food or technology literacy has been imbued in the general population. For example, all packaged food are required to have details of their ingredients etc. This may be one of the ways to ensure consumers of AI are similarly kept informed of what is done to their data. In broad terms we would support the proposed policy objective of AI operators having an obligation to explain in understandable terms how the AI in question works (without having to publish the algorithm itself). Also, we again refer to GDPR restrictions on automated decision making. Suggested alternative wording: "…any AI decision making which is likely to have significant impact on the rights and freedoms of individuals…"

rewording: "AI systems should not be designed in a way which intentionally enhances existing harms or creates new harms". We would also appreciate further detail from the AI HLEG on the specific harms envisaged in this context. AI HLEG text (The Principle of non-maleficence "do no harm" pp 9): To avoid harm, data collected and used for training of AI algorithms must be done in a way that avoids discrimination, manipulation, or negative profiling.Vodafone comments: Vodafone understands this could be quite an onerous requirement given the mainstream practice of profiling users to offer them more targeted services. Suggested rewording: "to avoid harm, data collected and used for training of AI algorithms must be done in a way which which avoids harmful/negative discrimination " and is in line with all applicable data protection and other laws, including the Charter on Fundamental Rights. Again, we are concerned that this principle lacks specificity: If something is intended to be prohibited or restricted, then the harms should be quite clearly articulated. There are unlawful (e.g. racial or gender based) and lawful discrimination (everything which is not prohibited by law - e.g. price discrimination). Is the intention to also limit the lawful discrimination? We ask the AI HLEG to provide additional clarity on lawful/unlawful discrimination in the context of the ethics guidelines. AI HLEG text (The Principle of Autonomy: "Preserve Human Agency" pp 9): If one is a consumer or user of an AI system this entails a right to decide to be subject to direct or indirect AI decision making, a right to knowledge of direct or indirect interaction with AI systems, a right to opt out and a right of withdrawal.Vodafone comments: The context of the use of AI is very relevant here. If AI is used within a communications network to improve energy efficiency or routing, there should be no need for a consumer to have a right to opt in or out of such use. However, if a consumer's data is being used to make decisions about that person using AI, they should be informed that this is happening. Where personal data is being used, an individual already has the ability to opt out under the GDPR so there is no need for additional requirements. However, GDPR does not make a distinction between "direct or indirect" interaction - but it refers to "legal effects" produced that affect the individual. This implies the direct or indirect distinction is irrelevant, it's the impact that matters. Vodafone would appreciate clarity from the AI HLEG on the interaction between the principle of autonomy as stipulated here, and the right to opt out of automated decision making as enshrined under GDPR, to assess how the two obligations would interact in practice. What we need is a clear culture or reading the data protection laws in a way that respects Human Rights and their fulfillment in data processing context.AI HLEG text (The Principle of Justice (be fair) pp 10): the positives and negatives resulting from AI should be evenly distributed,Vodafone comments: Vodafone questions the extent to which this is a realistic and practical requirement to include within these ethics guidelines. While all private actors engaged in the development of AI recognize the need to ensure these products and services work for the betterment of society, the requirement for all positives to be 'evenly

					distributed' is vague. This sort of metaphysical confusion arises out of lack of clearly articulated harms to be avoided, something which we ask the AI HLEG to rectify urgently. In our view most of the metaphysical confusion arises from the fact that the distributions mentioned in the text and which form the basis "Be Fair" principle should be applied to are not specified in advance with sufficient clarity. Grounding AI ethics in a clear legal framework derived from the Charter of Fundamental Rights and pertaining legislation would prevent much of this confusion arising in the first place. AI HLEG text (Identification Without Consent pp 11): Noteworthy examples of a scalable AI identification technology are face recognition or other involuntary methods of identification using biometric data (i.e. lie detection, personality assessment through micro expressions, automatic voice detection).Vodafone comments: Vodafone supports strong consent requirements around the development and use of facial recognition technology. Google's updated AI ethical principles blog explains how they have recently withdrawn a number of AI facial recognition projects on ethical grounds - this is a sensitive area and we recommend a robust approach to upholding individual privacy.. When processing biometric data with the purpose of identifying individuals, GDPR clearly states these are sensitive personal data (Art 9) for which stricter rules are needed. We would be obliged in any case to obtain explicit consent for this type of data.		
Michael	Lindholm	Turku Science Park Oy	Thank you for the opportunity to comment the "Draft Ethics Guideline for Trustworthy AI" The matter you are addressing is of great importance, it can shape the future in many ways. My comments are the following and based on the text in the guideline: I find the subject so wide that I do think that it is good to start with general guidelines very fast. I do believe that is good to focus on general guidelines and from there work out details. For that reason I like to comment on some things	1. Human in the centre – Individuals should have greater control already now of how his/hers data is used and also block the personal data or at least be able to follow the path how the data has been shared or from which data source the personal data has come. This is not the case right now. This is also an important thing for trustworthy AI. Without knowing from which sources the personal data and the possible decisions made by e.g. AI-algorithms there is no way for the person to decide on if the conclusions are right or wrong or if something is missing. Page 7, section 3.2 opens this a bit and also the transparency mentioned on page 10, I wonder how EU can handle the existing large players like Tencent, Google, Facebook etc.	2. The idea of developing quality systems (mentioned in COM(2018) 795) is very good. An extension to e.g. an ISO quality system would give the benefit of external control and also internal and external audits. I believe that a rapid development of the existing quality systems would be an effective way of getting trustworthy AI implemented faster than by regulating. I do not say that regulating should not be developed, I just think that quality system development would be more effective and faster. Saying this, SME:s should be considered in these quality systems. An ISO implementation is quite expensive.	3. Page 11, section 5.2: Why do we need systems to tell users that AI is used? I would recommend that companies should have the information available for all users of how the individual data is used and from where it origins. The methods of the e.g. decision making should be available. The last two sentences in this section raises many questions: Humanity from which point of view? Humanity and value of being human has changed a lot the last 100 years, how do we know what humans want the next 100 years? Got a comment from one person who said that AI and technology should be kept away from certain activities like yoga, meditation etc. to which I commented that certainly not because technology can help people in various ways also in these fields. 4. Page 12 section 5.3: I believe optioning out makes the design ruins the idea of scoring e.g. if the scoring is used in health applications to the benefit of all. Opt-out could also mean that the scoring will be biased. I agree that a general scoring of the individuals is not the best way but scoring could mean so many other things. 5. Page 15 section 2: Integrity of data: Who is capable and has the resources to monitor this? I believe this is could be a part of a quality system. 6. Page 16 section 4: See my comment 5 7. Page 18 section 8: I would strongly recommend that AI algorithm updates as well as the AI-connected systems updates should be considered in the fallback plan.	I did not have the document with me when I entered these comments. Based on an e-mail I sent you earlier

							8. Page 26 section 8 “accuracy through…”: Who can add data that affect the results? How is the added data validated? Could a parallel solution be implemented to validate results?	
David	PEARCE	Universidad Politécnica de Madrid, Spain		Explainable versus explicable AI (p.10)Explainable may be preferable. Explication is a technical term in philosophy referring to the rational reconstruction of concepts; most closely associated with the work of R. Carnap. I think explanations are what people expect from any kind of decision support system.LAWS, sec 5.4. Please consider deleting the sentence “Note that, on the other hand, in an armed conflict LAWS can reduce collateral damage, e.g. saving selectively children. " First, it suggests a possible endorsement of the use of autonomous weapons which is not in keeping with the ethical considerations expressed elsewhere in the document. Second, it suggests that some groups (i.e. children) are more valuable than others, which contradicts the earlier claims that all human life is equally valuable. (We could extrapolate this to moral dilemmas involving e.g. autonomous vehicles.)Moral agencyPage 13 footnote 19 claims that there are no artificial moral agents. In a literal sense this might be true if one believes that moral agency requires consciousness. However, normative multi-agent systems have been studied for many years. One might also defend the view that an artificial agent behaves morally if it’s intentions and goals are in keeping with accepted values and the same behaviour pattern would be considered morally acceptable if the agent were human. Footnote 24 on p.15 describes autonomy levels that would seem to support (artificial) moral agency.Incidentally, for autonomous agents having sensing, planning and acting capabilities, we can and should be able to inspect its values, intentions and goals at any time. Unlike in the case of humans, we can actually inspect the assumptions and reasoning mechanisms leading to a certain decision and action that the agent takes. This suggests that, like the black box in an aircraft, the artificial agent should always be equipped with a memory function that provides and maintains a complete trace of its decision making process. This should be part of the accountability process.	Design for all (page 15 and page 25).This should be expressed more clearly. Surely what is meant is that all areas of society should benefit from AI technologies. But the first sentence reads as if every AI system should be designed for all kinds of users. But clearly an educational tool for young children, a Matlab style mathematical assistant, and a social robot that cares for the elderly, are each designed for a different groups and purposes, and this is obviously not a problem.Undecidable logicsFootnote 24, p.15 the phrase “second order or modal logics are undecidable” is not very convincing. First, undecidability is a technical term that will not be familiar to many readers. Many modal logics are decidable. And moreover the point is not well taken because even first order logic is undecidable. Nevertheless we have very effective theorem provers based on first order logic.	Assessment list for autonomous driving/moving.You ask for inputs for this topic. I recommend the report published by the German Ethik-Kommission:Automatisiertes und vernetztes Fahren, Bericht Juni 2017, Bundesministerium für Verkehr und digitale Infrastruktur. www.bmvi.de.It is a very well-written document that clearly sets out and elaborates on many of the most important ethical issues, including questions of responsibility and problems of moral dilemmas. It is not the last word on the topic but provides a very useful framework for further extension and discussion. It includes 20 suggested ethical rules for autonomous and connected vehicles. I believe these are compatible with the general approach you take on trustworthy AI and provide a good basis for extending your draft document. I haven’t seen an English translation of the report, but it is possible that by now one exists.	
Anonymous	Anonymous	Anonymous	When you talk about "Trustworthy AI" you point at trust in the business and public governance. I think Dieselgate has shown neatly that business cannot be trusted and needs to be clearly regulated. While they still may find loopholes, a conservative approach will allow for trust.	1. You do not explicitly mention the right to privacy. For all the requirements for freedom and non-discrimination to apply, you must ensure that the data fed into the system is already stripped of any identifying information (compare to Apple's map routing algorithm). Only that way it's possible to guarantee freedom and non-discrimination 2. While you talk about equality in terms of access, you do not talk about equality in terms of treatment of people. E.g. a car insurance planting a tracking chip and then providing tailored insurances to their customers based on their driving habits is not covered under your guidelines. An AI algorithm optimising for ticket prices based on when people purchase them is not			Overall, the guidelines are way too fluffy and do not protect humans enough. It should be clearly stated, that if ever in doubt, the benefit should lie with the human, not the AI or the organisation behind it, irrespective if government or corporation or anything else.

				covered either. These principles oppose the idea of society (a construct that trades personal benefit for communal benefit). When talking about identification without consent, you say detecting fraud is aligned with ethical principles. Only if everybody agrees on what "do good" actually means. In case of rouge state operators, the lines between "fraud" and ethical principles may quickly become less clear. Which brings us back to utilitarianism of "is it ok to kill one person to save 10?" and where we draw the line? This is all very vague.			
nicolas	blanc	CFE CGC	In addition to playing a regulatory role,unions can be considered as addressees.	All relevant stakeholders that develop, deploy or use AI – companies, organisations, researchers, public services, institutions, unions, individuals or other entities – are addressees 3. Fundamental Rights of Human Beings 3.5 Workers'rights Workers and their representatives should enjoy a right to be informed of any automated treatment of their data by companies, and systematically be offered to express opt out (the same for people in hr recruitment). Workers (and people in hr recruitment) and should never be subject to systematic scoring by companies.	This also entails a responsibility for companies to identify from the very beginning the social impact that an AI system can have, and the social and legal rules that the system should comply with. Social-by-design is a new concept where developers, deployers and users : - are aware of social impacts on employment that an AI system can have - have to respect social legislation, local labor code that an AI system interacts with. Stakeholder and social dialogue : From better healthcare to safer transport, the benefits of AI are many and Europe needs to ensure that they are available to all Europeans. This requires an open discussion and the involvement of social partners, stakeholders and general public. Many organisations already rely on panels of stakeholders to discuss the use of AI and data analytics. These panels include different experts and stakeholders: legal experts, technical experts, ethicists, union representatives, etc. Actively seeking participation and dialogue on use and impact of AI supports the evaluation and review of results and approaches, and the discussion of complex cases Each year, all new or update automated treatments of the datas of the employees should be presented in employee representative committees (document SIA = Social Impact Assessment, Register of processing IA operations) KEY GUIDANCE FOR REALIZING TRUSTWORTHY AI Strive to facilitate the auditability of AI systems, particularly in critical contexts or situations. To the extent possible, design your system to enable tracing individual decisions to your various inputs; data, pre-trained models, etc. Moreover, define explanation methods of the AI system with a social-by-design method. Ensure a specific process for accountability governance in companies. Foresee training and education, and ensure that managers, developers, users and employers and their representatives are aware of, and trained in, Trustworthy AI.	7. Respect for Social rights Is the system Social compliant? For example, are there human micro taskings inside the treatment? How many ? What for ?(digital labor) Is the data information flow in the system under control and compliant with existing Labor rights? For example, interaction of the treatment (high frequency, night hours time) with human workers ? How can union representatives seek/find information about impacts on employment of the IA system? How can developers seek information about impacts on employment of their system? Is it clear, and is it clearly communicated, to whom or to what group issues related to social violation can be raised, especially when these are raised by users of, or others affected by, the AI system? Representatives of the employees ? Unions?	As stakeholder, union representative(s) should be included in this consultation, especially on Non-Technical Methods , to achieve Trustworthy AI.
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential

Anonymous	Anonymous	Anonymous	Trust in AI includes in first plays trust in the people deploying AI. There is no such thing as an ethical or trustworthy technology, there is an ethical an trustworthy deployment of technology, only. I think the HLEG should emphasize this, maybe by saying Trustworthy usage or deployment of AI and not Trustworthy AI.		There is no such thing as degree of autonomy. The Draft is mixing it up with the degrees of automatization. Or do you ever heard of peoples' different degrees of autonomy? It is important to emphasize that technology, in terms of a technical and legal perspective, not in terms of ontology, is about to become autonomous, wich means that it acts on the own (auto) rules (nomy)-	Your list with questions is one reasen for stifferin innovation. Please don't make the management of undertakings think, they might be using something that is under company law more alarming than the development of, lets say, cars with complex emission control. The managements need to know that while some framework conditions might change regarding AI, the overriding organisation principles of companies will still be true and stick.	Thank you a lot for giving so much thought to the subject. You are doing crucial work.
Gert-Helge	Dr. Geitner	TU Dresden, Germany					To whom it may concern, you added a list of members of the "High-Level Expert Group on AI". I am very interested in to know which members of this list took part in "Asilomar Konferenz 2017" and also who signed the guidelines of this conference? Was this document (guidelines of Asilomar conference) of any major interest regarding discussions of this Expert Group? I ask this question because there is only a footnote "6" on page 8. Are members of the expert group also members of the Future of life Institute? Thank you for your reply in advance, sincerely Gert-Helge Geitner
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Anonymous	Anonymous	Anonymous	Well written.	Well written.	Well written.	Well written.	Dear all, I would like to give some general feedback, which I believe is of utmost importance. I think each of the sections were very well written, but I did not see a general public section in reference to AI and emergency situations (please do correct me if I am wrong) i.e. access to personal information via object's AI devices and terrorism or other illicit criminal activity aka "the times when the ethics, principles and rights have to be ignored". This particular section , I believe, would be targeting the phone companies and mobile providers the most and then the users under suspicion. I believe "suspicions protocols" ought to be legally permitted in order to gather data directly from an object's phone screen, mic, audio, video without the object's concent and without the option for deletion of the sensitive data, even if the user has deleted it from the device, a black box type of alternative. The live access will be anyway limited I guess, but there should be an option for a black box software installation for people for whom a "suspicions protocol" has been initiated. This ought to be clearly outlined, fed to the public and expected. I think this way, the integrity of the user and the companies remains respected, (in the sense that they have been officially warned) but, also, security is taken into consideration and post factum, the entire package of data could be used to locate culprits, prevent other security threats, limit more damage. I strongly believe, we ought to openly discuss this when discussing AI. I believe this could be viewed as a detailed back up of legal surveillance, to be inacted, viewed and used only with special cases, of course. Often this theme is a "taboo", but it's a necessary theme to discuss directly considering the extremist activity in our modern world. I

						know the mentions of the law have been made throughout the text, but a more detailed assessment pertaining to this side of security would be even more beneficial. Also, I think, due to the , unfortunately, huge lack of education amongst the mass, the text, even though, it is so well written, must be edited, by adding examples in the text of what AI is ,as for example, stating the following: artificial intelligence meaning *definition*, which we use in our daily lives in the form of smartphones, smart cars and drones, social media feeds, music and media steaming services, video games, online ads networks, navigation and travel, banking and finance, smart home devices, security and surveillance and others. It's appalling, but people rarely comprehend the concept of AI, unless you directly tell them what it is to them specifically, in their daily life, with simple words and very clear examples. Otherwise, I think the text and graphs were amazing and interesting to read! The ideas for AI campaigns online and on tv, on the streets, in the offices, etc., for me, are endless. It's a limitless field full of creativity, but also a very responsible matter to execute as a job! As much as AI can help society, it can hamper it by being in the wrong human hands and directed for the wrong use. The laws, directives for artificial intelligence are directly connected to the local police authorities and laws, even though artificial intelligence and its respective laws transcend the local law, in a way. Developers and users share an equal responsibility to not abuse the rights, guidelines, laws and directives, but we often know that almost no one reads the rules and regulations before ticking the "I accept" box, so we would need to make sure this somewhat gray area is also taken into consideration when outlining the way AI ought to function, be limited to, etc. I strongly believe that if more people were aware of the fact that their Apple phone, for example, is AI , they would be more active in reading on the matter, giving feedback, etc. By knowing an AI device has their finger print and face biometrics, the same biometrics taken for their passport ID, I guess that should make them be more interested in the sovereignty they give up to a device and after all, to another person or people who are behind the AI device, be it as hardware or as software. I hope this was helpful. Thank you for the opportunity! I wish you success! Best regards,Avgustina Asenova Peycheva
Toby	Walsh	TU Berlin and . UNSW Sydney	I welcome the focus on ethical obligations that flow from human rights. However, it is important to note that this is necessary but not at all sufficient to ensure ethical AI. As an example, one area that this focus on human rights does not adequately cover is the actions of corporations. Human rights, by its focus on the individual, has little to say about how corporations act. However, we have some . strong expectations in Europe about how corporations should act -- we disapprove of monopolies, we expect corporates to act responsibly towards environment, etc. More fundamentally, human rights frameworks set a lower bound on good	5.4 Lethal Autonomous Weapon Systems (LAWS) It is highly disputable that "LAWS can reduce collateral damage". At most, you can say, "LAWS could reduce collateral damage". We have little evidence to show that smarter weapons (since, for instance, the Gulf War) have reduced collateral damage. Indeed, LAWS are the perfect weapons of terror, that will be used by rogue states and terrorists to target civilian populations, so have the potential to increase the collateral impact of warfare. The report identifies some critical concerns (Sec 5). One area not discussed that I believe should be is the use by governments of AI decision support systems in high stake	This chapter would be improved by considering the international dimension of how to realise trustworthy AI. One of my hats is Chair of the Expert Working Group preparing an AI plan for Australia, and member of group drafting ethical guidelines for the use of AI. Tech companies act globally. We need then to act globally, to exploit international fora (ISO, IEEE, D8, UN ...) and international rules and frameworks to ensure all of us on the planet benefit from AI. With respect to this, I was disappointed not to see ANY mention of the global south. It took many years and much action (some ongoing) to see, for instance, that the pharmaceutical		

			behaviours. And we can, and should expect AI to be well above these lower bounds,	areas that impact on citizens' rights (e.g. provision of welfare, sentencing, etc.)	industry acted ethically in its relations with the developing world. I fear unless we keep this region in mind, they will miss out and be exploited in the development of AI.	
Alexandra	Smyth	Royal Academy of Engineering, BCS - the Chartered Institute for IT and Institute of Measurement and Control	A cross-engineering sector response to the European Commission's High-Level Expert Group on Artificial Intelligence Draft Ethics Guideline for Trustworthy AI on behalf of the following UK organisations:Royal Academy of EngineeringBCS, the Chartered Institute for ITInstitute of Measurement and Control.Purpose and scope of the guidelines• The ethos of the guidelines is, for the most part, admirable. Ensuring that AI has a human-centric approach is a sensible top-level aim and building trust in AI as a technology is an important part of achieving this. However, there needs to be further clarification of what is meant by 'human-centric'. Specifically, it would be useful to include examples of non-human-centric AI systems in order to clarify the definition.• The guidelines ought to be principally a framework for trust and ethics within AI currently and in the short-medium term future rather than attempting to forecast into the distant future. Developments should continue to be monitored until greater agreement of the future concerns become apparent. • A number of points apply to all computing or digital systems and are not specific to the subset of these systems that use AI technology. Distinguishing which aspects of the document apply to all digital systems - for example, the fundamental rights of human beings - and which apply specifically because the computer system is built on AI technologies would make the report more impactful. Clarity about what distinguishes AI systems from non-AI systems and how the applicability of ethical principles and values vary accordingly would be useful.• There are tensions between rapid innovation and ethics; for example, carrying out detailed evaluation of risks and testing of new systems that incorporate AI technologies may be in tension with the requirement to be first to market. Notwithstanding, ethical concerns can enhance innovation as well as restrain it - a trusted service is very evidently more popular. The document should foster innovation in AI rather than restraining it, and not be too prescriptive for an industry that is still in its infancy. • It will be important to inculcate principles on AI on a global level, not just within the EU, for them to reach their full potential. • The document is currently academic in nature and the challenge will be to translate these principles into useable guidance for practitioners.• The main omission in these guidelines is any discussion of the need for better public understanding of AI. Recent successful AI applications have provided World-class performance in a very narrow area. Such 'intelligence' is outside the experience of most people. There is a danger that they will overestimate the capabilities of such systems by trusting them beyond their capacity. Additionally, unrealistic expectations about the capabilities of AI systems can arise without an appropriate understanding of how they work. • The importance of education, both for	• A number of the principles may make sense at a high-level, but may be problematic when applied to specific AI technologies or in specific contexts.• For example, at a high-level it is desirable to develop AI technology that does no harm but AI technology is already used, and will continue to be developed, in weapons systems.• The 'equal distribution of economic, social and political opportunity' from AI is also challenging in practice. While few people would disagree that the benefits of AI should be spread equitably across society, the distribution will depend on the national economic, political and social contexts in which AI is being applied. • Building ethical principles in AI from existing treaties and charters, such as the EU Charter of Fundamental Rights, is a sensible approach from both a legal and ethical basis. Although, in some instances, the document infers more from existing treaties and charters than is clearly present. For example in section 3.4, the document puts forward that equality in an AI context entails "a fair distribution of the value added being generated by [AI] technologies". However, the EU Charter chapter on equality only refers to equality before the law and in terms of discrimination against protected characteristics, not in terms of distribution of resources. Consequently, care needs to be taken that when the document argues for something in an AI context, it is clear where principles derive directly from rights and where they derive from ethical theories or other sources.5. Critical concerns raised by AI• As society becomes increasingly dependent on socio-technical capabilities incorporating AI, the possibility of failure of such a capability and its potential impact are increasingly critical concerns. • A key concern is that the guidance is based on a singular 'AI system', which is engineered and operated by one organisation which has authority over design, developmental and operational aspects of the system. This self-containment would enable the responsibilities and expectations placed on the technical, commercial, legal, ethical aspects of the 'AI system' to be monitored. The concept of a neatly self-contained 'system' is an anachronism in today's world of increasingly interconnected and inter-dependent systems, invariably developed and operated by different organisations, with different stakeholders, motivations and objectives, subject to not entirely consistent rules and regulations. Services incorporating AI capabilities are evolving organically and continually, driven by the interests of multiple stakeholders.There is a lack of robust, scalable methods for practical implementation of the principles for trustworthy AI. Existing technical and non-technical methods may possibly be applicable to small, closed systems but do not seem fully adequate for the engineering and governance of the interconnected and inter-dependent systems already being constructed and deployed. Conventional	1. Requirements of Trustworthy AI: 2. Data Governance:• Data must be assembled, structured and managed over its lifecycle so that it meets requirements, for which a robust engineering approach is needed. This will help to provide assurances about data quality, provenance and timeliness. Good quality metadata is vital. Considerations such as whether the data is being updated, or how it might be securely destroyed, are also relevant here, as are the security of data storage and transmission. People with the necessary data curation skills are needed to manage the data. If data is shared between organisations, it will be vital to ensure that data is being managed in line with data-sharing agreements between parties exchanging and using the data.• An important part of oversight is being clear which aspects of the system or decisions that the human wishes to be responsible for and for which ones will they pass over responsibility to the AI system.8. Robustness: • The definition of 'robustness' appears to be a catch-all for a range of often distinct elements. There are at least two subsections that could be distinguished. The first, 'robustness and security' concerns the potential failure/exposure of the AI system under attack or failure. The second, 'reliability and accuracy', concerns the effectiveness of the AI system to achieve its function/intention over a range of inputs and circumstances. This is also related to the description of function/intention and the concept of traceability. The 'robustness' definition here seems non-standard and should be reconsidered in all its varieties.9. Safety• In describing 'safety' there is a conflation between function/intention and safety. Here it is stated that 'Safety is about ensuring that the system will indeed do what it is supposed to do…' but this description is not safety but the intended function of the system. Safety should primarily be concerned with 'whatever the system does, it should not harm users, resources or the environment'. The question of what it does, or intends to do, is irrelevant here. The definition of safety should be re-cast so that questions about the intended function of the system are moved into other sections, such as 'transparency'. If an AI system is to be transparent then it should be clear what it is supposed to do and what it is trying to do (function and intention).• There may be difficult trade-offs between safety, security and privacy.10. Transparency• It is crucial that there is transparency about the function and intention of an AI system, or in other words what it is trying to do and how it is trying to achieve this. 2. Technical and Non-Technical Methods to achieve Trustworthy AI1. Technical methodsArchitectures for Trustworthy AI: • This is fundamental and the point that 'the requirements should be integrated at 'sense'-level' is key.Testing and validating: • As the document acknowledges, conventional testing approaches are inadequate. Statistical sampling is often inappropriate as the behaviour of an AI is	• Much of the practical implementation of Trustworthy AI is still in the research stage. Companies are attempting to implement regimes that comply with many of these requirements, but the approaches used are under continuous refinement. 1. Accountability• Accountability presents a number of mechanisms to help reach its aims. One suggestion is responsible AI training which should be a priority. A requirement for all employees in AI to engage in some sort of AI Ethics course would help to improve knowledge and skills on the subject. One mechanism could be to introduce an ethics module in higher education courses likely to produce graduates who will work in AI. At a lower level, standardised staff training in AI Ethics, of the sort that is commonly seen for other tech issues such as data protection, could be a method to ensure that knowledge of these issues is not just the purview of developers and experts. • The idea of establishing a sustainable mechanism for oversight, such as an internal or external review board, is vital. The question of how organisations make decisions about 'grey areas' in AI is a challenge. The review board should ideally represent a diverse mix of expertise as well as employees, partners and customers. Mechanisms for sharing best practice are required.7. Respect for (& Enhancement of) Human Autonomy:• Section 7 on human autonomy does not make clear what implementing its recommendations would look like or how its recommendations would be delivered. Highlighting every algorithm and assuming the user has the knowledge to make an informed decision on its use does not seem realistic and would probably end up being similar to how many people click terms and conditions boxes indifferently, despite ostensibly being presented with the information that should help to make it a more informed decision. • As AI develops, there will be machines using data in ways we have not thought of yet and that are incredibly complex to understand. Consequently, the idea of 'full self-determination' in decision making put forward in this section is unrealistic and a more helpful approach would be to help people reach a base level of understanding. This reiterates the need for education and awareness from school onwards about what AI is and how it operates in general and relatively accessible terms. Only by providing people with this basic information can things like interrogating algorithms or providing information on AI products become intelligible and genuinely increase human autonomy in this area.8. Robustness• Robustness could use a more holistic approach with other technologies and be expanded as a result. For example. AI is a component of a wide range of Internet of Things (IoT) products and the entirely of these need to be robust and resilient to attack if we are to reduce the chance of them being compromised. • Good data management is a vital part of achieving	A cross-engineering sector response to the European Commission's High-Level Expert Group on Artificial Intelligence Draft Ethics Guideline for Trustworthy AI on behalf of the following UK organisations:Royal Academy of EngineeringBCS, the Chartered Institute for ITInstitute of Measurement and Control.

practitioners and users of AI, cannot be overstated. Many of the recommendations within section III (Assessing Trustworthy AI) are predicated on a level of AI understanding not currently present. Some combination of basic teaching to pupils at schools and ethical components within relevant higher education qualifications would help to address this situation.

systems engineering approaches, characterised by the 'V Model' (https://incoseonline.org.uk/Documents/Groups/Engineering_and_Project_Management/SEPM_V_Model_for_ASEC2104.pdf), are unsuited for engineering these capabilities and services. New methods will be necessary to enable the design, development, operation and ongoing maintenance of such systems of systems.• In the UK, research programmes such as the Assuring Autonomy International Programme (https://www.york.ac.uk/assuring-autonomy/) are addressing the lack by technical and non-technical methods, by developing suitable standards, guidance, and technical methods and toolsets, along with educational and training resources.5.1 Identification without Consent• Practical tools are required to help citizens understand explicitly the trade-offs when they acquire 'free' applications or functionality in exchange for use of their personal data. In practice, much current 'consent' is not 'informed'. • More work is required to investigate the extent to which a database of 'anonymised' data can be queried before it is effectively 're-personalised'.• Identification without consent needs to be regulated, as appropriate to the intent. 5.2 Covert AI systems• The onus should be on deployers of such technology to make it clear whether the user is interacting with a human or AI, but some work would be needed on a taxonomy that would facilitate public understanding.5.3 Normative & Mass Citizen Scoring without consent in deviation of Fundamental Rights• It seems essential to allow opt-out but again people need to have the consequences explained to them.5.4 Lethal Autonomous Weapon Systems• While technologies such as Lethal Autonomous Weapons (LAWS) may be incongruent with the ethical guidelines put forward by the draft, the reality is that the development of this technology is already happening. LAWS will proceed regardless of any codes of AI Ethics. Instead specific international conventions will be needed to limit this. The United Nations Group of Government Experts (GGE) on LAWS is also looking at the ethical dimension of this technology . • It is important to recognise that all weapons are subject to International Humanitarian Law (IHL), including those with high levels of automation and autonomous weapons if developed. IHL places strict constraints on their design, control and use through the Geneva Conventions and their Additional Protocols. All states are subject to IHL so it is essential that autonomous weapons remain under state control for design and use. Ethical concerns arise if non-state actors use them for aims which are not widely considered to be ethical and used outside IHL constraints.• The definition of LAWS vary. Banning LAWS now based on definitions of an AWS responding to a commander's intent, which is highly futuristic, may allow unrestricted development of applications of AI and other high levels of automation in weapon systems that are currently not allowed under IHL. This was generally agreed, including by the EU delegation, at the 2018 meetings of the UN Group of Governmental Experts (GGE) on LAWS. It is the reason why the UN did not request an immediate ban despite pressure from lobby groups.• The draft states: 'in an armed conflict LAWS can reduce collateral

typically discontinuous near (semantic) boundaries in the input domain, a fact exploited by many published successful attempts to 'fool' AI systems. • As well as testing, which is quite a weak verification technique, there are other forms of verification that can be used. Specifically, formal verification whereby we 'prove' some behaviour of a system will always/never occur is important here. This stronger form of verification is crucial if we are to have strong guarantees of behaviour. Such formal techniques are important both in the verification of decision-making and the verification of intention (when the AI system is responsible for some key aspect). Throughout this subsection it should be clear that there are a range of verification techniques of varying strengths and that it will be crucial to use stronger techniques where the greater risks occur.• Notwithstanding, formal verification methods are not a panacea. In the case of neural nets, verifying that a neural net tool, which decides on the weights and carries out the back-propagation functions, satisfies a specification is feasible. However, these verification methods do not apply to the classification application which uses the tool to train a neural network on large datasets to be able to classify inputs correctly. The application's correctness depends on a range of variables including the representativeness and size of the training set. Biased training sets will lead to incorrect behaviour even using a verified neural net tool. Overfitting will occur if there too many features are used given the size of the training set. Traceability and auditability• As well as tracing/explaining decisions it is important to trace the intent of the AI system. If it is to make a decision or act on its own, then what it is aiming to achieve should be transparent at all times. Explanation (XAI research)• The requirement for trustworthiness states: 'Explainability – as a form of transparency – entails the capability to describe, inspect and reproduce the mechanisms through which AI systems make decisions …'. But in many statistical machine learning systems, such explainability is not obtainable: their mechanisms are inherently opaque. This is briefly alluded to on p20 'A known issue with learning systems based on neural nets is the difficulty to provide clear reasons for the interpretations and decisions of the system'. However, elsewhere the guidelines refer to, for example, 'the causality of the algorithmic decision-making process' and 'AI systems should document both the decisions they make and the whole process that yielded the decisions, to make decisions traceable', even though such traces are not always obtainable.• Actuarial decisions have long been based on statistical data, without requiring causal explanations of the mechanisms by which particular factor(s) affect the likelihood of a risk materialising. Data analytics facilitates the 'discovery' of (previously unexpected) statistical associations or correlations within (large) datasets; whilst these associations may be statistically robust and safe to exploit in certain (non-critical) circumstances, the causal relationship may remain elusive. In many respects, Machine Learning is merely an extension of these statistical analyses and inferences – and has similar limitations. The decisions reached by machine learning AI take into consideration tens of factors

accuracy through data usage and control. Data must be assembled, structured and managed over its lifecycle so that it meets business or other requirements, for which a robust engineering approach is needed. Considerations such as whether the data is being updated, or how it might be securely destroyed, are relevant here.9. Safety• There is a tendency to wait until something goes wrong, either accidental of maliciously (e.g. Gatwick drones), before we take corrective action. Every effort should be expended to anticipate and mitigate these risks at the outset.10. Transparency• Ensuring transparency of algorithmic decision-making is a challenge, particularly for machine learning and self-adaptive systems. Issues of governance and accountability will need to be considered in the design and development of these systems so that incorrect assumptions about the behaviour of users – or designers – are avoided. • Transparency of the data on which the algorithmic decisions are being made is critical to ensure accountability. But with more transparency there are additional risks to markets, privacy, disincentivising companies from developing IP or people gaming the system. These trade-offs need to be understood in the specific context of their application and therefore more nuance is needed.• Section 10 in relation to transparency seems difficult to implement as it stands, due to the breadth of its definition in the context of the draft guidelines. The definition of transparency used requires certainty over product benefits, usage scenarios and product limitations and this is not realistic, especially for a technology that is still in its early stages and may be used by many operators or customers.

damage’. This makes unrealistic assumptions about the capabilities of current autonomous systems to distinguish, for instance, combatants from non-combatants, especially when informal forces are involved. Failure to make such distinctions breaches the Geneva Convention.• There are ethical concerns about the increased use of AI in the decision-aids used by commanders in making weapon-release decisions, for example target-identification. Target and collateral object recognition systems using AI and machine learning require training in realistic scenarios. Unlike civilian applications, there are very few or no actual combat scenarios suitable for training machine-learning systems. This gives concerns about the reliability of their results even if one does not consider that many people and objects in the scenario will be actively deceiving their opponents using sophisticated techniques.Additional critical concernsThere are additional critical concerns such as:• The use of AI for manipulating democratic systems.• The use of chatbots engaging with children and young adults, perhaps shaping their views in ways which if identified would raise concern.• The use of AI for manipulating financial markets as we move to a time when more sophisticated control is possible whereby a hostile agent learns to manipulate the behaviour of market to enable it to cause major disruption at will.• The setting of realistic expectations from AI systems. For example, with AI health systems there can be many benefits, however, machine learned classifiers tend to ignore outliers so those with an unusual pathology may be missed. The total death rate may be reduced but the death of just one individual through a machine error will raise ethical questions and issues of liability.• How responsibility is allocated is an important consideration. For example, is it ethical for an autonomous car to be less safe than an autonomous aircraft just because it is accepted with human drivers.5.5 Potential longer-term concerns• Active cross-disciplinary collaboration between, for example, computer scientists and neuroscientists is helping to push the state of the art, and narrow AI is beginning to apply lessons learned from one environment to another. However, one of the central challenges in achieving general AI is ‘transfer learning’ – the ability of computers to infer what might work in a given scenario based on knowledge gained in an apparently unrelated scenario - which is not something they currently can do. Although the timescale for general AI goes well beyond the next 20 years, it is critical to understand now how to solve the ‘control problem’ as it is such an important issue. • Notwithstanding the importance of being aware of potential longer-term concerns, the guidelines ought to be principally a framework for trust and ethics within AI currently, and in areas where there is a general consensus on the short-medium term future. Dealing with this alone is already a significant subject matter and attempting to forecast what will happen at a later juncture, especially when there is little agreement among the authors, does not seem necessary. The guidelines consequently should not look to deal with uncertain situations a long way off at this point, but continue to monitor developments until greater agreement develops.

making them very difficult to interpret.• Additionally, XAI (Explicable AI) will either increasingly hamper the development of new AI-based capabilities, some of which could be highly beneficial to humankind. The ‘explanations’ may be incomprehensible to (most) human minds. Alternatively, the AI system may be expected to produce an explanation in text, which will be an approximation due to the complexity. This is neither useful nor open to effective scrutiny or oversight.• Notwithstanding, not all machine learning algorithms are black boxes. It is also possible to use alternative ways of achieving explainability, such as counterfactual reasoning.Other technical methods:• Human factors design: A particular system design issue is how best to give the operator the right information to exercise appropriate control. Human factors design of the system is key to achieving human oversight - the ability of the human to easily and correctly assimilate what the AI is doing to enable appropriate intervention, if necessary. In complex systems the operator may need to be highly trained to deal with decisions handed over by the AI to the human. • Accident investigation: It is essential to be able to investigate accidents and incidents. Access to this data will be valuable but opens up questions about commercial sensitivity. • Monitoring: Continuous monitoring of a capability may help to delay or lessen the impact of failure, but the inevitable trade-offs between false positives and false negatives could limit its usefulness.2. Non-technical methodsRegulation• Regulation requires a level of consistency across sectors and applications to help achieve good governance, public understanding and confidence. However, while certain principles may apply across all sectors and applications, regulation will need to be developed on a sector-by-sector basis, taking into account the criticality of the application and the existing regulatory context. This approach should acknowledgethe disparate requirements and constraints of sectors or domains along with their points of commonality. Common approaches across sectors and domains will also help to avoid duplication and support multi-sector supply chains and applications. Sectors must work together to develop common approaches, and also to ensure consistency between policies for existing cross-sectorial applications or those that may emerge in the future.Ethics and inclusive design education• Ethics education is needed at all stages of the pipeline, for schoolchildren and adults. There is a role for professional institutions in education and training, and to build on existing ethical frameworks developed for individual members of a profession, such as the UK engineering profession's Statement of Ethical Principles which forms the basis for codes of conduct. However, one challenge is that not all people who work in the field of AI associate themselves with a particular profession or are members of a professional institution.• There is also a need for diversity and inclusion to be a part of education, and for guidance on inclusive design principles for AI.Practical implementation of the guidelines by organisations• Tensions may exist between an organisation's values and the values of its individual workers. Ethical principles of individual employees need to be

					reconciled with an organisation’s brand values and expectations of the people it serves. It will be important to co-create ethical frameworks with both employees and employers, and to engage with customers and partners on an on-going basis to ensure ethical concerns are addressed. • Ethical and inclusive design principles need to be part of existing processes in organisations such as governance and quality assurance. There needs to be an internal accountability framework, and a way of auditing any AI system. Quality assurance processes would help to ensure that an algorithm has been adequately tested and is sufficiently accurate, that the data used in training the algorithm and in subsequent analysis is appropriate and that checks have been made to assess the data for bias, for example. Incentives would be of benefit, such as including ethics within key performance indicators. It will also be important to build ethics into the roles of those involved in developing and using algorithms. • To implement the ethical guidelines within organisations will be challenging but necessary to develop ways of assessing whether ethical requirements have been met.		
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Thilo	Weichert	Netzwerk Datenschutz expertise	(1) Das Ziel, auf EU-Ebene gemeinsame ethische Grundlagen zu erarbeiten, die bei der Erforschung, der Entwicklung und dem Einsatz von Künstlicher Intelligenz (KI) zu beachten sind, ist zu begrüßen.(2) Zu kritisieren ist, dass sich die Leitlinien auf KI beschränken, also auf informationstechnische Systeme, die aus Sensoren oder anderen Quellen stammende Daten aufbereiten und hieraus „selbstlernend“ ursprünglich von Menschen gestaltete Algorithmen verändern und auf dieser Grundlage automatisierte Schlussfolgerungen bzw. Ergebnisse gewinnen, die zur Grundlage von relevanten praktischen Entscheidungen genommen werden (können). (3) KI ist eine Weiterentwicklung von hochkomplexen Algorithmen. Bei komplexen Algorithmen, die nicht auf selbstlernenden, sondern auf vorgegebenen ausdifferenzierten Datenauswertungsprozessen beruhen, bestehen ähnliche Herausforderungen, wie sie von der HLEG bzgl. KI in den Guidelines thematisiert werden. So werfen z. B. nicht auf KI-Basis funktionierende Scoring-Verfahren ethische Probleme fehlender Verantwortlichkeit, Zurechenbarkeit, Transparenz und Kontrollierbarkeit auf.(4) Das Thema von ethischen Leitlinien sollte	(10) Die Leitlinien benennen richtig als ethische Vorgaben die Grundrechte, insbesondere die Menschenwürde, die Freiheits- und Bürgerrechte, die Diskriminierungsverbote, sowie die Grundsätze von Demokratie, Rechtsstaatlichkeit und Solidarität (S. 7). Diese Grundsätze haben ihre verfassungsrechtliche Grundlage in der seit 2009 wirksamen europäischen Grundrechte-Charta (GRCh) gefunden. Nicht thematisiert wird die weitergehende Frage, inwieweit es durch die Digitalisierung einer Weiterentwicklung der verfassungsrechtlichen Normierung bedarf. Der insofern gestartete Prozess der Formulierung digitaler Grundrechte (https://digitalcharta.eu/) muss in den weiteren Diskussionen über die vorliegenden Guidelines ein zentraler Aspekt sein.(11) Im Rahmen dieser verfassungsrechtlichen Diskussion bedarf es der Erörterung, inwieweit das Grundrecht auf Datenschutz, das vom deutschen Bundesverfassungsgericht (BVerfG) als „Recht auf informationelle Selbstbestimmung“ definiert wurde (BVerfG 15.12.1983 – 1 BvR 209/83 u. a.), um ein „Recht auf digitale Souveränität“ zu ergänzen ist, das auch juristischen Personen	(13) Entgegen einer weit verbreiteten Wahrnehmung dient die DSGVO nicht nur dem Schutz des Grundrechts auf Datenschutz, sondern dem Schutz „aller Grundrechte und Grundfreiheiten natürlicher Personen“ bei der personenbeziehbaren Datenverarbeitung (Art. 1 Abs. 1 DSGVO). Die Guidelines reduzieren die Anwendung der DSGVO auf den Respekt von Privatheit (Privacy, S. 17, 25). Dadurch wird auch ignoriert, dass die DSGVO sämtliche relevanten Bewertungskriterien vertrauenswürdiger KI einer Regulierung zuführt: Verantwortlichkeit, Design, Selbstbestimmung, die Verhinderung von Diskriminierung, Robustheit und Richtigkeit, Sicherheit und Transparenz (S. 24-27). (14) Die DSGVO thematisiert umfassend den Grundrechtsschutz von Betroffenen bei personenbeziehbarer Datenverarbeitung sowie die damit verbundenen gesellschaftlichen Konsequenzen. Grundrechtsrelevante Wirkungen entfalten sich nicht nur bei der Verarbeitung personenbezogener Daten, sondern auch, wenn automatisierte Entscheidungen ganze Personenkollektive betreffen und hierbei sächliche oder vollständig anonymisierte Daten verarbeitet werden. In der weiteren Diskussion müssen anwendungs- und	(15) Mit der DSGVO besteht bisher schon ein verbindlicher gesetzlicher Rahmen für den Einsatz von KI in Bezug auf Profiling und automatisierte Entscheidungen mit personenbeziehbaren Daten. In Art. 22 DSGVO werden Abwägungsanforderungen benannt, die bei der Gestaltung, dem Einsatz und der Nutzung von KI einfließen müssen: individuelle Selbstbestimmung (Einwilligung), Eingreif- und Revisionsmöglichkeit (z. B. Ausschaltknopf), Ersetzungsmöglichkeit durch einen menschlichen Entscheider, besonderer Schutz beim Einsatz sensitiver Daten, Rechtsschutz). In Art. 15 Abs. 1 lit. h DSGVO wird das Recht auf Auskunft über „die involvierte Logik sowie die Tragweite und die angestrebte Auswirkungen“ begründet.(16) Ein weitergehender Rechtsrahmens für den Einsatz von Algorithmen im Allgemeinen und KI im Speziellen sollte daher auf diesen bestehenden Normen aufbauen. Die DSGVO gibt hierfür den notwendigen Spielraum (vgl. Art. 22 Abs. 2 lit. b DSGVO). Damit wird zugleich gewährleistet, dass weitere verfassungsrechtliche Anforderungen, die in den Guidelines nicht oder nur andeutungsweise erwähnt werden, beachtet werden. Dies gilt insbesondere für den	

daher nicht auf „künstliche Intelligenz" beschränkt werden, sondern generell den Einsatz komplexer Algorithmen umfassen. Zielsetzung der EU sollte es demnach sein, generell einen normativen Rahmen für den Algorithmeneinsatz und die Algorithmenkontrolle zu definieren.(5) Bei datengetriebener KI besteht wegen des dauernden Prozesses der Selbstoptimierung durch Datenauswertung und der Nachjustierung der Entscheidungsfindung eine noch geringere Bestimm- und Bachvollziehbarkeit als bei determinierten digitalen Prozessen. Dadurch verschärfen sich die generellen Probleme automatisierter Entscheidungen in Bezug auf Protokollierung, Transparenz, Verantwortlichkeit und Haftung. Diskriminierungseffekte können nicht nur durch die Programmierung bewirkt werden, sondern solche Effekte werden durch einfließende Daten aus realer Diskriminierung verstärkt. Eine Dokumentation und Nachvollziehbarkeit der Entscheidungsfindung ist nicht mehr gewährleistet. Eine individuelle Verantwortung für Einzelentscheidungen wird vorverlagert von der Systemgestaltung hin zur Entscheidung über das „Ob" eines Systemeinsatzes.(6) Daher bedarf es hinsichtlich des KI-Einsatzes weitergehender Restriktionen oder Vorkehrungen. Für bestimmte Zwecke ist der Einsatz von KI-Technologie wegen der damit verbundenen Konsequenzen überhaupt nicht ethisch vertretbar und muss deshalb absolut ausgeschlossen werden. Dies gilt z. B. für den militärischen KI-Einsatz bei tödlichen Waffen; dies gilt aber auch bei nicht-militärischen Nutzungen, wenn die per KI getroffenen Entscheidungen existenzielle Bedeutung für Menschen haben und keine Revidier- bzw. Kompensierbarkeit besteht. (7) So wichtig ethische Standards sind, so bleiben diese unverbindlich, wenn sie nicht in bestimmte Gesetze oder sonstiges zwingendes Regelungen umgesetzt werden, die demokratisch zustande gekommen sind, deren Einhaltung unabhängig kontrolliert und deren Verletzung effektiv sanktioniert wird. Dieser Prozess der Operationalisierung der Leitlinien wird in den vorliegenden Guidelines nicht thematisiert. Ohne diese Operationalisierung besteht die Gefahr, dass den ethischen Leitlinien ein reiner Alibicharakter zukommt und dass diese zur Legitimation für ethisch problematische Techniknutzungen eingesetzt werden.(8) Nicht nur der Prozess der Normsetzung wird in den Guidelines übergangen, sondern weitgehend auch die Relevanz von Normen generell: Um vertrauenswürdige KI zu erlangen, wird in erster Linie auf technische Methoden gesetzt. Dabei wird zutreffend differenziert zwischen Technikgestaltung, Architekturen, Testung, Bewertung, Dokumentation und Erklärbarkeit (S. 19 f.). Zu kurz kommen die Prozesse der regelmäßigen Kontrolle und Evaluation, die bei KI als Systemen, die auf lernenden, also sich ändernden Algorithmen basieren, besonders wichtig sind.(9) Hinsichtlich der nicht-technischen Methoden wird auf Standarisierung, Governance, Verhaltensregeln, Erziehung und auf einen gesellschaftlichen pluralen Diskurs Bezug genommen (S. 21 f.). Diese Methoden sind zu ergänzen durch eine unabhängige Zertifizierung (s. u. Rn. 18) und eine unabhängige menschliche Kontrolle (s. u. Rn. 17). Die Notwendigkeit demokratisch zusteht und nicht nur für von digitaler Verarbeitung Betroffene gilt, sondern auch für solche Techniken (verantwortlich) Anwendende (also Nutzende). Digitale Souveränität ist ein Ziel, das nicht nur für die Objekte von Datenverarbeitung (also Betroffene im datenschutzrechtlichen Sinn) realisiert werden muss, sondern auch für die Systemnutzenden als Subjekte. Ein zentrales Problem des Einsatzes künstlicher Intelligenz besteht darin, dass die diese (verantwortlich) Nutzenden auf die Technikbereitstellung durch Anbieter angewiesen sind, deren Angebot sie weder bewerten und einschätzen, geschweige denn verantworten können. Die Idee wird in den Guidelines nur angedeutet (S. 9 f.).(12) Das Prinzip der Erklärbarkeit bzw. der Transparenz von KI ist ein Grundanliegen der Guidelines (erstmals S. 10, dann z. B. S. 18). Dieses Prinzip ist eine Grundvoraussetzung nicht nur für KI, sondern für digitale Datenverarbeitung generell. Diese Prinzip ist auch grundlegend für die Realisierung des Grundrechts auf Datenschutz (Art. 8 GRCh) und ein zentrales Anliegen der dieses Grundrecht umsetzenden Europäischen Datenschutzgrundverordnung (DSGVO, dort z. B. Art. 5 Abs. 1 lit. a, 12 ff.).

zweckbezogen die Bereiche identifiziert werden, in denen derartige Anwendungen eine derartige Relevanz entwickeln, dass regulative Ergänzungen zu den bestehenden Regelungen zur Verarbeitung personenbezogener Daten nötig sind (z. B. in den Bereichen Mobilität, Umweltschutz, Nahrungsmitteleinsatz, Biotechnologieeinsatz).

Auskunftsanspruch der Betroffenen bzw. in einem erweiterten Verständnis der Anwendenden als „digitalen Souveräne" (s. o. Rn. 11) sowie für die unabhängige staatliche Kontrolle (Art. 8 Abs. 2 S. 1 u. Abs. 3 GRCh).(17) Die Notwendigkeit einer unabhängigen staatlichen Kontrolle wurde beim Datenschutz schon früh vom deutschen Bundesverfassungsgericht (BVerfG) verfassungsrechtlich begründet, insbesondere für den Einsatz digitaler Technik durch staatliche Einrichtungen (erstmals BVerfG 15.12.1983 – 1 BvR 209/83 u. a.). Sie wurde vom Europäischen Gerichtshof (EuGH) mehrfach eingefordert (EuGH 09.03.2010 – C-203/15 u. C-698/15, 16.10.2012, - C-614/10, 08.04.2014 – C-288/12). Die diese Rechtsprechung tragenden Erwägungen lassen sich auf die Kontrolle von KI generell im öffentlichen wie im privaten Bereich übertragen.(18) In der DSGVO ist in den Art. 42 f. der rechtliche Rahmen für die Zertifizierung komplexer informationstechnischer Systeme durch eine freiwillige unabhängige Überprüfung festgelegt. Die hierfür nötigen Instrumente müssen umgehend in der Realität umgesetzt und angewendet werden. (19) Für den grundwertekonformen Einsatz von KI genügt in vielen Bereichen eine freiwillige Zertifizierung nicht. Es bedarf, wie beim Technikeinsatz in anderen gesellschaftlichen Bereichen üblich (z. B. bei der Mobilität, bei Emissionen, beim Gentechnikeinsatz, bei Arzneimitteln) einer darüber hinausgehenden bereichsspezifischen Regulierung mit Melde-, Genehmigungs- und Evaluationspflichten und einer einsprechenden hoheitlichen Kontrolle. (20) Aus der Diskussion in den USA kommend, wird auch in Europa teilweise die Position vertreten, von Algorithmen errechnete Ergebnisse könnten den Schutz der Meinungsfreiheit (Art. 11 GRCh) für sich in Anspruch nehmen. Diese Argumentation wird eingesetzt, um eine stärkere Regulierung von KI bzw. eine verstärkte Algorithmenkontrolle zurückzuweisen. Es ist notwendig, in den Guidelines klarzustellen, dass die Nutzung von Ergebnissen digitaler Datenverarbeitung, insbesondere von KI, für sich nicht das Grundrecht auf Meinungsfreiheit in Anspruch nehmen kann.(21) Die Guidelines vermeiden bei dem Ziel der Herstellung vertrauenswürdiger KI bzw. generell von vertrauenswürdigen digitalen Entscheidungsprozessen eine Aussage zu einer grundlegenden Fragestellung: Von Verantwortlichen wird dem Transparenzerfordernis der Schutz von Betriebs- und Geschäftsgeheimnissen entgegengesetzt. Tatsächlich hat z. B. das oberste deutsche Zivilgericht, der Bundesgerichtshof, entschieden, dass Betriebs- und Geschäftsgeheimnisse selbst Transparenzforderungen an digitale Prozesse entgegen gehalten werden können, die von datenschutzrechtlich Betroffenen geltend gemacht werden (BGH 28.01.2014 – VI ZR 156/13, BGH 22.02.2011 – VI ZR 120/10). Berechtigte Verfassungsklagen hierzu wurden bisher vom deutschen BVerfG nicht angenommen (dazu Weichert DatenschutzNachrichten 2/2018, 134). Der EuGH hat sich mit dieser Problematik bisher nicht befasst.(22) Ein zentrales Problem beim Einsatzes von KI ist, dass die in den Guidelines aufgeführten ethischen Werte bei vielen konkreten KI-Einsätzen in der Praxis nicht beachtet werden, weil die diese Verfahren einsetzenden Unternehmen, bei

getroffener Regeln bzw. Gesetze wird nicht ausdrücklich, sondern nur in sehr allgemeiner Form thematisiert (S. 21). Tatsächlich ist ein klarer, mit Verboten und Geboten, technisch-organisatorischen Vorgaben und prozeduralen Regeln festgelegter gesetzlicher Rahmen, dessen effektive Einhaltung gewährleistet wird, die zentrale Grundlage eines vertrauenswürdigen Einsatzes komplexer Algorithmen.

denen es sich sehr oft um mächtige Wirtschaftsunternehmen aus den USA wie Google, Facebook, Amazon oder Microsoft handelt, sich bisher erfolgreich weigern, die zur Umsetzung der Grundrechte und der demokratischen Kontrolle nötige Transparenz herzustellen. Durch diese Transparenzverweigerung, für die angeblich bestehende ökonomische Rechte ins Feld geführt werden, wird eine wirksame Rechtskontrolle unmöglich gemacht. Um dieses zentrale Problem in den Griff zu bekommen, bedarf es klarer gesetzlicher Offenlegungspflichten der einsetzenden Unternehmen gegenüber der demokratischen Öffentlichkeit bzw. gegenüber staatlichen Stellen sowie einer hinreichenden Ausstattung der unabhängigen Aufsicht, damit diese Pflichten auch praktisch durchgesetzt werden können. Eine europäische Regulierung ist wegen der europa-, ja weltweiten Bedeutung des Problems wünschenswert. Demokratie- und Grundrechtskonformität muss Vorrang haben vor Marktverfügbarkeit, Wettbewerb und ökonomischem Nutzen.(23) Durch eine Vorverlagerung des Risikos beim KI-Einsatz von der Gestaltung des Einsatzes digitaler Technik hin zur Entscheidung, ob diese eingesetzt wird, muss im Sinne eine Gefährdungshaftung zumindest eine zivil- und verwaltungsrechtliche Verantwortlichkeit gesetzlich begründet werden. Die KI einsetzenden Stellen müssen per Gesetz spezifischen Gestaltungs- und Unterlassungspflichten sowie einer umfassenden Haftung unterworfen werden.

laetitia	Pouliquen	Page 2: We would like to stress that the following phrase seems either inaccurate: “The goal of AI ethics is to identify how AI can advance or raise concerns to the good life of individuals, whether this be in terms of quality of life, mental autonomy or freedom to live in a democratic society. It concerns itself with issues of diversity and inclusion (with regards to training data and the ends to which AI serves) as well as issues of distributive justice (who will benefit from AI and who will not).”One definition of ethics by Merriam Webster is: “the discipline dealing with what is good, bad, with moral duty and obligation; a set of moral principles; a theory or system of moral values”. [1] the question at hand is not to define what is the “good life of individuals” but rather set the line between good and bad which is a far larger objective than what the HLEG expresses and includes individuals with “bad” quality of life such as the mentally and physically challenged citizens. We believe ethics should not give a clear-cut decision on what a “good life” is. As reminded by the Oviedo convention, fundamental rights are the foundation to ensure the “primacy of the human being” in a context of technological evolution.We propose that this sentence read as follows: “The goal of AI ethics is to identify how AI can advance or raise concerns to the primacy of human beings over technology, to ensure respect for human rights such as freedom of being who they are by virtue of being humans. This leads to the ethical principle of autonomy which prescribes that individuals are free to make choices about their own lives, be it about their physical, emotional or mental wellbeing (i.e. since humans are valuable, they should be free to make choices about their own lives). In turn, informed consent is a value needed to operationalize the principle of autonomy in practice. Informed consent requires that individuals are given autonomy to live in a democratic society. It concerns itself with issues of diversity and inclusion (with regards to training data and the ends to which AI serves) as well as issues of distributive justice (who will benefit from AI and who will not).”In Purpose and Target Audience of the Guidelines page 2"A mechanism will be put in place that enables all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis. This will be set out in the final version of the document".This point is important. Precision will be necessary to specify the obligatory value of the Guidelines. If it is not obligatory, there is a risk that companies may find that taking ethics into account will undermine innovation and profitability. It will also be necessary to provide a means to value the companies that approve the Guidelines.	In 3. Fundamental Rights of Human Beingspage 7“At the same time, citizens should enjoy a right to be informed of any automated treatment of their data by government bodies, and systematically be offered to express opt out”How do citizens access services when they opt out? How they still be served by AI based companies when they decide to refuse the algorithm-based decision-making process? In comparison, user experience after the adoption of GDPR, certain websites totally block their access to their content upon refusing cookies. Even if technical cookies are accepted by the user, some companies still refuse to serve the user with its goods and services. We shall see the same bottle neck effect when users reluctant to AI-based decision will not be able to perform their commercial activities as well as their administrative tasks. the expulsion of some users will be inevitable and will de facto undermine the notion of freedom of choice and autonomy. In 4. Ethical Principles in the Context of AI and Correlating Values page 8“It should also be noted that, in particular situations, tensions may arise between the principles when considered from the point of view of an individual compared with the point of view of society, and vice versa. There is no set way to deal with such trade-offs. In such contexts, it may however help to return to the principles and overarching values and rights protected by the EU Treaties and Charter. Given the potential of unknown and unintended consequences of AI, the presence of an internal and external (ethical) expert is advised to accompany the design, development and deployment of AI. Such expert could also raise further awareness of the unique ethical issues that may arise in the coming years.”This paragraph is unclear: which experts? Reporting to whom? Could you please clarify on what unlikely tensions you refer to?In 5.3 Normative & Mass Citizen Scoring without consent in deviation of Fundamental Rights Page 12“We value the freedom and autonomy of all citizens. Normative citizen scoring (e.g., general assessment of “moral personality” or “ethical integrity”) in all aspects and on a large scale by public authorities endangers these values, especially when used not i situations, tensions may arise between the principles when considered from the point of view of an individual compared with the point of view of society, and vice versa. There is no set way to deal with such trade-offs. In such contexts, it may however help to return to the principles and overarching values and rights protected by n accordance with fundamental rights, or when used disproportionately and without a delineated and communicated legitimate purpose. […] However, whenever citizen scoring is applied in a limited social domain, a fully transparent procedure should be available to citizens, providing them with information on the process, purpose and methodology of the scoring, and ideally providing them with the possibility to optout of the scoring mechanism.”First, why a “limited social domain”? Algorithmic scoring and notation impact very large areas that are not by definition limited: they are bound to have a "snowball" effect. Credit scoring impacts access to housing, to employment, to proper schooling etc.Second, as already mentioned, by excluding oneself from the AI-based	In 1. Requirements of Trustworthy AI 1. AccountabilityPage 14“Good AI governance should include accountability mechanisms, which could be very diverse in choice depending on the goals. Mechanisms can range from monetary compensation (no-fault insurance) to fault finding, to reconciliation without monetary compensations. The choice of accountability mechanisms may also depend on the nature and weight of the activity, as well as the level of autonomy at play. An instance in which a system misreads a medicine claim and wrongly decides not to reimburse may be compensated for with money. In a case of discrimination, however, an explanation and apology might be at least as important”.This is a very important point. The "accountability" must be distinguished from “liability” (ie. to be legally responsible). Point 1 should establish the distinction between accountability and liability.In . Requirements of Trustworthy AI 4. Governance of AI Autonomy (Human oversight)Page 15 We have two remarks :“The level of autonomy results from the use case and the degree of sophistication needed for a task”. This is agreed upon. We recommend an idea by Researcher Cyrille Dalmont in Intelligence artificielle et santé : 10 propositions anti - brouillard pour régulation éclairée. The user should be informed on the level of AI used to reach a decision by allowing for the identification of diagnoses and prognosis made by artificial intelligences. For this purpose, a pictogram could be affixed on any document, image or prescription produced by an AI. The patient would then be able to identify the degree of human involvement in the conclusions made to the medical examinations carried out and have a recourse if need be.Similarly, autonomy of decision lies in the anonymity of data. Cyrille Dalmont proposes : “The collection and processing of patient data is a crucial issue. risk could simply be color-coded based on the degree of confidentiality or sensitivity of the data and their treatment with state-level labeling of companies and their level of entitlement to process certain data according to precise specifications and security guarantees provided by authorized public companies and organizations. By way of illustration, the data enabling predictive medicine to be carried out should be classified as the most sensitive with an absolute ban on dissemination to certain institutions or economic actors such as insurance companies, banks or lessors in order to avoid Digital precariousness. An individual could no longer get access to insurance, medical treatment, contract a loan or rent a home if his/her risk factors were too important.»Second remark : "FOOTNOTE 24. AI systems often operate with some degree of autonomy, typically classified into 5 levels: (1) Domain model is implicitly implemented and part of the programme code. No intelligence implemented, interaction is based on stimulus-response basis. Responsibility for behaviour lies with the developer. (2) Machine can learn and adapt but works on implemented/ given domain model; responsibility has to be with the developer since basic assumptions are hard coded. (3) Machine correlates internal domain model with sensory perception & information. Behaviour is data driven with regard to a mission. Ethical behaviour can be modelled	Page 28The HLGE solicits our partaking in the practical operationalization of the assessment list on four particular use cases of AI, selected based on the input from the 52 AI HLEG experts and the members of the European AI Alliance: (1) Healthcare Diagnose and Treatment, (2) Autonomous Driving/Moving, (3) Insurance Premiums and (4) Profiling and law enforcement. (1) Healthcare Diagnose and Treatment Bioethics requires special vigilance and attention to the announcement of the "bad news". If the diagnosis and prognosis have been established by an RN, care must be taken to ensure that doctors pay more attention to the way information is presented, based on the psychology of the patient and the humanity of the caregiver. (2) Autonomous Driving/movingAutonomous cars communicate with each other: car companies must ensure that shared data remains private and anonymous.In addition, transportation laws differ from member state to member state. It seems necessary to work on the harmonization of transport rules or at the very least on the road signing so that autonomous cars can operate optimally and without damage risks.(4)Profiling and Law Enforcement The burden of proof is a part of the rule of law in the EU. AI based law decisions must not invert the burden of proof principle and place it on the user, should the AI-based law decision be contested.	As a preamble, we would like to acknowledge the quality of the draft produced by the HLEG on AI and to reckon that we are off to a very good start. This draft seems to be a powerful basis to which we, Professor Nathalie Nevejans, Robotics and Artificial Intelligence Law and Ethics Expert, Lecturer in Law, University of Artois (France), Member of the CNRS Ethics Committee, Expert to the European Parliament and Laetitia Pouliquen, Director, NBIC Ethics, are honored to contribute.First, we are truly grateful that our recommendation on Ethics to the European Commission for more transparency on the impact assessment progress of the Machine Directive 2006/42/CE was taken into account. This impact assessment leading to possible changes in regulations on the critical issue of machines and algorithms liability is now possible via the Machinery Directive revision feedback until next February. Second, the make-up of the HLGE of 52 experts remains unbalanced: with an overwhelming number of industry and federations stakeholders, we stress out the rare number or absence of philosophers, ethicists, religious leaders, anthropologists, consumer organizations and health experts. An enhanced AI HLEG would better guarantee the necessary respect for human rights based on a deep human-machine understanding. Please refer to our previous post with our Ethical recommendations on AI published in December 2018. Third, it is noteworthy that the Oviedo Convention was not adopted by all members of the Council of Europe. Moreover, even among the signatories, several nations and states took a very long time before ratifying the Convention. Consequently, the question we ask is “Will these AI ethical guidelines turn into soft law to ensure the cultural shift needed for AI developers to take onboard these ethical constraints?”Fourth, another thought-provoking question remains: why should we ask for a willing base adoption when human rights are de facto overarching rights? Should we sign on rights that are already defined in the EU Charter of Fundamental Rights and the European Convention on Human Rights, be it for AI or anything else? However, to ponder this remark, the technical variations of the EU ethical principles are novel and need to be designed.Fifth, ‘Trustworthy AI’ could also be branded as ‘Ethical Inside’ (like the famous and efficient ‘Intel Inside’) with an ISO accreditation.

algorithms, he/she shall not benefit from what he or she claims, subject to the evaluation of the algorithm (eg: a bank credit).Please add facial recognition for commercial use as strictly forbidden for “ethical integrity”. In 5.5 Potential longer-term concernsPage 12 ADD Human-Machine responsibilityWe note that the Joint Research Center report - Artificial Intelligence: A European Perspective JRC rightly recalls the principles and EU values impacted by AI : autonomy, identity of individuals, dignity and right to privacy and personal data protection. The JRC paragraph on dignity draws our attention to the possible erosion of human rights: “Individuals’ rights and responsibilities could start eroding as a result of the increasing interaction of humans and machines (EDPS, 2018). At the moment, smart devices have no moral responsibility and that is why it could be potentially harmful to let them manage human beings (EGE, 2018). However, the European Parliament called for the EC to consider a specific legal status for robots (EP, 2017), which is still a controversial proposal when considering, for instance, that at the present time accountability is ultimately related to human responsibility (EECS, 2016).”First, it is inconceivable that individuals’ rights and responsibilities should erode as a result of the increasing interaction between humans and machines. Second, the fact that, “at the moment, accountability is ultimately related to human responsibility” (EESC) is a good thing. However, we reiterate that the creation of a specific legal status for robots would be the wrong response the liability problem, as expressed in our Open Letter to the European Commission on AI and Robotics. Signed by 285 EU experts in AI, ethics and law, the signatories hereby affirm that the creation of a Legal Status of an “electronic person” for “autonomous”, “unpredictable” and “self-learning” robots is inappropriate from a technical, ethical and legal perspective. Humans must always be responsible for their algorithms and for any damages caused. In fact, The European Group of Ethics of Science and Technology denies any moral standing to AI systems or robots in its report from March 2018 Artificial Intelligence, Robotics and ‘Autonomous’ Systems . EGE shares its moral reflections against the principle of autonomy. AIs are algorithms and robots are machines: “Human beings ought to be able to determine which values are served by technology, what is morally relevant and which final goals and conceptions of the good are worthy to be pursued. This cannot be left to machines, no matter how powerful they are. […] Moral responsibility, in whatever sense, cannot be allocated or shifted to ‘autonomous’ technology.”[2]Similarly, UNESCO’s World Commission on the Ethics of Scientific Knowledge and Technology (COMEST) confirms their ban on a legal status for AIs and robots in its report of COMEST on robotics ethics calls the possible creation of a legal status for robots as: “highly counterintuitive to call them ‘persons’ as long as they do not possess some additional qualities typically associated with human persons, such as freedom of will, intentionality, self-consciousness, moral agency or a sense of personal identity”[3] Page 12 ADD - Holistic view over AI AI and according to decision logic with a utility function. (4) Machine operates on a world model as perceived by sensors. Some degree of self-awareness could be created for stability and resilience; might be extended to act based on a deontic ethical model. (5) Machine operates on a world model and has to understand rules & conventions in a given world fragment. Capability of full moral judgement requires higher order reasoning; however, second order or modal logics are undecidable. Thus, some form of legal framework and international conventions seem necessary and desirable. Systems that operate at level 4 can be said to have “Operational autonomy”. I.e., given a (set of) goals, the system can set its actions or plans."With regard to Footnote 24, the definition, and especially the legal consequences for autonomy, lack a nuance and finesse, and are very obscure concerning points 3 to 5. It is not possible to approach as succinctly the questions of civil liability on such subtle points of distinction. And above all, these words suggest that the HLEG on AI believes that in this case, it is someone other than the developer who should be responsible, that is to say the machine itself. It would therefore be necessary to delete the legal references or modify the text. Here is our proposal for deletion in Footnote 24:"24. AI systems often operate with some degree of autonomy, typically classified into 5 levels ; as autonomy increases, the determination of the responsible person may be more difficult: (1) Domain model is implicitly implemented and part of the programme code. No intelligence implemented, interaction is based on stimulus-response basis. (2) Machine can learn and adapt but works on implemented/ given domain model; (3) Machine correlates internal domain model with sensory perception & information. Behaviour is data driven with regard to a mission. Ethical behaviour can be modelled according to decision logic with a utility function. (4) Machine operates on a world model as perceived by sensors. Some degree of self-awareness could be created for stability and resilience; might be extended to act based on a deontic ethical model. (5) Machine operates on a world model and has to understand rules & conventions in a given world fragment. Capability of full moral judgement requires higher order reasoning, however, second order or modal logics are undecidable. At Levels 4, but especially 5, a legal framework and international conventions seem necessary and desirable. Systems that operate at level 4 can be said to have “Operational autonomy”. I.e., given a (set of) goals, the system can set its actions or plans." In 1. Requirements of Trustworthy AI 8. RobustnessPage 17 In matter of reliability and resilience to attack, we should mention the obligatory measure of a “kill-switch” button to AI automated robots. In 1. Requirements of Trustworthy AI 9. Safety Page 18 “Moreover, formal mechanisms are needed to measure and guide the adaptability of AI” .What does this exactly mean? In 2. 1 Technical methods "Ethics & Rule of law by design (X-by-design)" Page 19"This also entails a responsibility for companies to identify from the very beginning the ethical impact that an AI system can have, and the ethical and legal rules that the system should comply

Robotics are artefacts that could impact our humanity. Nevertheless, the line between human and machine must be unequivocally affirmed. Therefore, AI needs to be viewed holistically due to NBIC technology convergence: Nano, Bio, Information and Cognitive technologies will all use AI algorithms and interact with people, either externally or internally.Boundaries between restorative and augmentative health technologies Eg of NeurosciencesThe European Commission should set the line between restorative and augmentative technologies and decide whether augmenting human beings using AI is acceptable or not, with regards to EU values. We believe that promoting an augmented humankind would increase unfairness and inequality and lead to a loss in individual s' rights in a democratic EU society. As an example, let us pick the interaction between AI and neuroscience. Swiss researchers Marcello Ienca and Roberto Andorno [4] give a relevant example. Their research led them to believe that human rights in the age of neuroscience and neurotechnology are subject to four major threats: • Right to cognitive freedom as the right to alter one's mental state by technical means and the right to refuse to do so. It is in fact the right not to be pressured to reveal data.• Right to mental privacy as the right to prevent illegitimate access to our brain information - This is in fact the question of neuromarketing.• Right to mental integrity as the right of individuals to protect their mental dimension from any potential danger, for example from hacking by a neural device (hacking of a neuro-device) • Right to psychological continuity as the right to preserve one's personal identity and consistency of the individual behavior against unacceptable changes, even if the changes introduced are not per se dangerous.Their research illustrates how, as AI is now used in all neurological technology, its impact has the potential to challenge who we are as humans. AI and robotics need to be considered holistically and not just as one element of a more global use of NBIC technologies. Is allowing the augmentation of humankind with AI or else, acceptable according to the EU values?We still need to address these philosophical and anthropological questions: what are the boundaries between human intelligence and artificial intelligence? Between humans and physical machines? Between natural life and Artificial life? The more we know about AI, the more it calls for a profound reflection on the boundaries between human intelligence and artificial intelligence. Determining what defines us as human beings will avoid the blurring between natural life and artificial life. Without this reflection, the questions of EU values and human rights would be irrelevant.The lines between restorative care and augmentation of humans need to be set. The European Commission should decide whether augmenting humankind using NBIC technology is acceptable to EU values. Investing significant EU funds for human restorative care technologies is desirable. However, we believe that augmenting humankind would result in unfairness and inequality. Individual rights would be more difficult to guarantee, even in a democratic society.

with".We perfectly agree. Nevertheless, it should be noted that it is very difficult to identify all the ethical impacts "from the very beginning". If all ethical impacts cannot be identified immediately (due to the novelty of the product, for example), they must be able to be identified later (after an experience of the product on the market and/or its varieties of use, for example).In addition, this question raises the problem of whether only companies should be concerned. We believe that both States and the Public authorities should equally be concerned and responsible in the matter and should also be encouraged or obliged to endorse them."Explanation (XAI research)""[…] In addition, sometimes small changes in some values of the data might result in dramatic changes in the interpretation, leading the system to confuse a school bus with an ostrich for example. This specific issue might be used to deceive the system".It seems very relevant to give an example. There should be more in the guidelines.

Aida	Ponce Del Castillo	European Trade Union Institute	(1) Let's not repeat the mistake that was made with the 'green-washing' a few years ago, which cost a lot in term of reputation to the European Commission and possibly led to European citizens losing trust in the institutions. In other words, let's us not end up with an "ethical-washing" of AI. Instead, make sure we develop strong, enforceable and binding regulation-based and consistent rules/principles for all companies and authorities that can enforce them. (2) AI is the most disruptive technology that we had in several decades, workers are worried about their future and how that will impact their life and jobs and that of their children. This concern cannot just be overlooked or disregarded. Experts claim that with AI some jobs will disappear, but other jobs will be created, the problem with that is that with the new jobs that will come and those that will go away are not interchangeable. Re-skilling is not the solution and it is not going to work for everyone. Some sort of protective buffer needs to be embedded into the labour market globally (Ponce del Castillo, 2018). (3) Many voices in Europe claim that if we legislate Al then we will lack behind other world super powers (USA and China). Not legislating at all will not guarantee that we will win that race but will guarantee that we will guarantee an EU with a lot of social unrest and a more fragmented labour market.	The ethical purpose of the guidelines promotes a unilateral and voluntary industry-action. Moreover, it limits trade union action. A unilateral approach that relies only on industry is not the right way to do this. We need both sides to be involved : industry and unions as unions are the voice of workers. Respecting ethical values and principles is valuable but it is unlikely to be effective, because it will essentially rely on the industry's own incentives and there is no system to monitor oversight or to solve issues when values get in conflict with other values. Previous experience with other technologies (Observatory of Nanomaterials) or with the Supply Chain Initiative (SCI) in the food sector also suggests that purely voluntary initiatives are not suited for creating a functioning independent redress mechanism and fairness rules that are attractive and credible for both sides of the market.	The selection of the principles looks only to the side of the developers and there is no mention to the principles of precaution, prevention, solidarity, common good nor distributive justice. We need more than code of conducts, declaration of principles and 'private governance' mechanisms, because this is too important to be entrusted to developers, companies and innovators without a sanction system. It is too important to base it in code of conduct and principles and nobody will get punished if they are not respect it. Ethical principles are not associated with any sanction system. the relationship risk/reward is so unbalanced, that some actors may decide that it makes financial sense to break or disrespect the principles. Having followed the way nanotechnologies have been 'regulated' for the past 10 years, there are similarities in that process, and the AI 'regulatory' process taken place today. We need to cannot end up with a toothless Observatory of AI, like has been done with nanomaterials. We need proper legislation and an appropriate mechanism to monitor AI developments. Legislation that can be revised to adapt provision to AI related developments are: General Data Protection Legislation, Product Safety Directive, directive on Liability for Defective Products, Directive on Safety at work, and Medical Devices Regulation.	What is important is to implement technology through monitoring mechanisms, so values are effectively respected. Being judge and party at the same time does not work. The whole responsibility cannot be dumped the on developing ethical AI on to developers, there needs to be an external eye looking at this and a way to attribute liability. Let's avoid a situation in when ask developers only to ensure that AI is developed in a 'ethical' manner. Nobody should be allowed to be judge and party instead, we need an external monitoring and confirmation that indeed the principles are respected. "Minimum regulatory standards need to be developed in order to attribute responsibility and liability in cases where the artificial agent has 'learning and teaching' features and is able to exercise unintended outcomes" (Ponce del Castillo, 2017). An effective regulatory framework is ultimately required in order to ensure that artificial agents co-exist harmoniously with humans and that they are specifically designed for, operate according to and are capable of adapting to human values and needs. Regulators will need to figure out how to manage risks and attribute liability, particularly as machines increasingly acquire the ability to learn and take independent decisions. Without a legal framework, transparency and trust will not exist, which will be detrimental to everyone, even the industry. (Ponce del Castillo, 2018).	
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Philips			Right to decide to be subject to AI/right to opt-out/ right of withdrawalPages 10 - 11 include a right to either be subject to AI/a right to opt out and/or a right of withdrawal. A right to decide to be subject (or not) to AI, a right to opt out and a right to withdraw significantly reduces the possibility to make use of AI systems. By definition, AI relies on large volumes of retrospective data, making the execution of these rights impossible for any AI system, especially since typically AI systems will further use the input by users to improve the algorithms the AI system is built of. In addition, these requirements were not omitted in the GDPR. On the contrary, GDPR, which regulates data protection, provides already for very specific requirements regarding automated decision-making. As an example, page 09-10 stipulates:"If one is a consumer or user of an AI system this entails a right to decide to be subject to direct or indirect AI decision making, a right to knowledge of direct or indirect interaction with AI systems, a right to opt out and a right of withdrawal".We propose to limit this sentence to the following:"If one is a consumer or user of an AI system, this entails a right – at any time during the use – to decide to be subject to direct or indirect automated decision making, a right to knowledge of direct or indirect interaction with AI systems."In addition, page 12, paragraph 5.3 includes a right to opt-out from any scoring mechanism: "and ideally providing them with the possibility to opt-out of the scoring	Page 15 includes an unclear statement on data governance which should be deleted, namely this one: "To trust the data gathering process, it must be ensured that such data will not be used against the individuals who provided the data." The AI guidelines should not lead to a situation where, for instanc,e patients whose personal data was used for development of an AI system detecting cancer cells, cannot profit from the future use of that AI system to have their own cells checked. Page 15 also includes a"design for all" requirement. This paragraph needs to be clarified, to reflect that AI systems can be designed for specific user groups and need to be (only) user centric for the targeted user group. For instance, AI systems intended for use by medical specialists do not need to be tailored to the lay knowledge of the average individual, as the systems will only be used by medical specialists.	All the comments made above also apply to Chapter III (the assessment). More in particular, the above comments should be specifically reflected in the following items: 3. "Design for all":o "Is the system equitable in use?o "Does the system accommodate a wide range of individual preferences and abilities? "6. " Respect for Privacy" o "How can users seek information about valid consent and how can such consent be revoked"	Philips welcomes this public consultation as it gives stakeholders, who are not members of the HLG on AI, the opportunity to provide inputs and comments on the draft guidelines. However, according to Philips, some aspects/notions/sentences would need to be amended/clarified or deleted, as indicated in the sections above.

		mechanism" and also states: "Developers and deployers should therefore ensure such opt-out option of the technology's design, and make the necessary resources available for this purpose." It is difficult to see how to comply with such a requirement, as data will be interwoven with the algorithm. Therefore we propose that this is limited to situations in which consent is the legal basis for the personal data processing by the AI system. If not consent, but another legal basis is used, for instance legitimate interest, there should be no requirement with regards to opt-out functions.Pages 10 and 11 refer to informed consent. It is not clear whether the document prescribes informed consent as a hard requirement for any AI system. The assessment in Chapter III (see box below) seems to indicate that this is the case. Basing the data processing for AI exclusively on informed consent will seriously hamper the use of AI, as it leverages large volumes of retrospective data. We believe that the GDPR safeguards provide sufficient protection for any AI system, as AI is a specific form of data processing. GDPR already ensures a legal basis, transparency, explicability, human intervention in automated decision making and accountability. However, GDPR identifies six legal basis for processing personal data, of which consent is only one. The question is, why would we confine the legal basis to 'consent' only?			
German Aerospace Center (DLR; EU Transparency Register No. 21280626733-05), Executive Board Representative Digitalisation	We, the German Aerospace Center (DLR), appreciate the human-centred approach chosen by the High-Level Group. We agree that technical robustness, as well as ethical purpose, must be the basis for a trustworthy AI. In its research and innovation projects on AI, the DLR builds on trustful systems. The DLR therefore welcomes the coordinated long-term plan for Trustworthy Artificial Intelligence "made in Europe," in addition to the work of the high-level expert group on AI. It is encouraging that the working group is planning and proposing a continuous process of necessary discussion. This addresses the challenge, that this document does not (and cannot) cover all conceivable cases. Furthermore, we take a very positive view of the envisaged possibility for stakeholders to voluntarily commit themselves to this guideline. This might lead to a gradual commitment by more and more companies, organisations, and researchers to support the goal of AI systems that are useful to people.	We fully support most of the statements and explanations in this chapter. The section "Do no harm" requires AI to be environmentally friendly, thereby mentioning the earth's natural resources. From our point of view, this requirement extends to space, where AI is used already today (e.g. on the international space station and for satellite-based applications). We support the statements in the section, "The Principle of Explicability." In our view, explicability also entails accountability for the quality and fair choice of training data. In our opinion, the fifth section on critical concerns is essential to the discussion. Here, we propose to add another concern. It is currently becoming clear that AI is being used as a method of research. In the sense of good scientific practice, research results must still be reproducible and verifiable in the future. If AI is used as a method of research, the AI used and data must be published, or at the very least, a means of testing for requirements needs to be established. Otherwise, peer review procedures would become more difficult (if not impossible), and a two-class research landscape might emerge. We say a two-class research landscape, as certain organisations that conduct research using AI systems may generate results which would no longer be reproducible. The brief mention of potential long-term concerns is appropriate. However, consensus building in this area is likely to be more time-consuming due to the numerous unclear future developments in the field of AI. We therefore propose that these issues are addressed in a separate process to	The guidelines for the implementation of trustworthy AI seem to us very useful. This applies to the requirements, as well the technical and non-technical methods. In the second and fifth requirements it is mentioned that bias should be removed from data. To us this seems to be a more or less impossible task. It is more realistic to recommend reducing rather than removing bias, and to call for increased awareness of the omnipresence of bias. Regarding the "Testing and Validating" method, we would like to add that the problem with data in the context of Open Data is becoming even more critical. Open Data means that the data does not originate from controlled or even known sources. The use of such data sets as a starting point for critical AI systems is problematic, as the traceability of its integrity is essential. In general, cybersecurity becomes more and more important for IT developments. The DLR sees cybersecurity as a crucial concern for all of its research applications. Thus, we recommend a closer link between cybersecurity measures and AI, especially when it comes to regulation and standardisation. The European Union has, or plans to have, useful cybersecurity measures in place (e.g. the certification framework for cybersecurity or the permanent mandate of the EU cybersecurity agency ENISA). This is also closely linked to user aspects since new products should follow a certification procedure that ensures strict security standards. Such certification could also apply for clear standards on transparency and accountability of learning systems, including legal obligations for producers of such products and services. We therefore see the	As a first draft, the list of questions is very adequate. However, there is a risk of stakeholders not feeling addressed by the way the questions are structured. The "User Stories" method could be helpful, since it always integrates a role. A possible template could look like this: As a "ROLE," do I know … ? For example: "As an AI developer, do I know what measures have been taken to ensure that the AI system always makes decisions that are under the overall responsibility of human beings?" At a later stage, it might be useful to address end users of AI products and applications with a similar set of questions.	We, as a neutral advisor acting on scientific insights, welcome and support the development of ethical guidelines on AI. As a research organisation that maintains a strong focus on (and long experience with) the development of critical high-tech infrastructures and systems, including verification of their trustworthiness, we will endeavour to follow and help shape such guidelines. We thank the High-Level Group for their contributions to date. In order to ensure the transparency of the consultation process, we ask to take note of our EU Transparency Register ID: 21280626733-05.

				develop dedicated guidelines on the long-term concerns and establish an ethical technology monitoring process. In medical research, the consultation of an ethics commission is mandatory. This might also be a viable solution to certain research on AI. Maybe the definition of criterion could help facilitate the evaluation of cases, where an ethics commission should be consulted, prior to the onset of the project.	topic of cybersecurity relevant in sections II 1.8, II 1.9 and II 2.2. At the end of the chapter, the question of further technical and non-technical methods is asked. There might be some approaches from agile software development, as a non-technical method, that could be worth mentioning. In particular, some methods and approaches from this agile context seem to fit very well with the approaches recommended in the document, including: "pair programming" as a method to identify potential threats and ethical conflicts during the development of an AI system, "continuous integration" and "short feedback loops" as methods to continuously compare the behaviour and goals of an AI with our goals, as well as "refactoring" as an adapted method to adjust the behaviour of an AI system according to feedback, etc. We would like to supplement that for the development, implementation and use of AI in critical systems such as autonomous driving, as the application of these agile methods alone is not sufficient to achieve the desired level of security and safety. This will require the definition of new methods and processes to verify AI-based software.	
ENRICO	CAIANI	European Society of Cardiology - WG eCardiology - Advocacy committee	We agree that AI needs to be human centric. In particular, if applied to health, this implies even higher responsibility, with the patient benefit and risk reduction at the center of attention. The declared aim at page 1 ("to foster a climate most favourable to AI's beneficial innovation and uptake") cannot be reached, in particular in the healthcare domain, without complete transparency relevant to results of validation studies in support of the use of AI and describing clearly potentials and limitations [Fraser A et al, The need for transparency of clinical evidence for medical devices in Europe. The Lancet 2018;392(10146):521-530. doi:10.1016/S0140-6736(18)31270-4].In fact, trustworthy AI in healthcare needs an additional component (page 2): to show clinical value and effective benefit for the patient, not just working as expected ("technically robust and reliable"). As regards target audience, it would be better to include in the discussion main stakeholders, instead than asking for formal endorsement once the Guidelines will be finalized. In fact, in multiple parts of the document there is reference to AI applications in healthcare, but no healthcare organization is listed among the 52 members of the High-Level Expert Group on Artificial Intelligence, as well as patient involvement appears limited (only a member of the board of the Austrian Association Supporting the Blind and Visually Impaired is in the list).Not including the end users (medical professionals and patients) as main participants in the process could severely hamper the correct uptake of this technology, thus replicating errors of the past in the healthcare field when new technologies were introduced to the market by emphatic marketing, but with no real clinical value and sometimes with potential damage for the patient.Due to the need of a tailored approach (see page 3) where ethical principles have to be entangled with specific daily problems and scenarios, maybe a solution would be to create a specific High-level Expert Group, where representatives of different medical associations and patients are included, in order to properly decline the ethical problems in practical scenarios, and having the health of the patient as only compass to find the right solution.	The organisation and delivery of health services and medical care is a responsibility of each member state in the EU. Nonetheless, the Treaty of Lisbon of 2007 sanctioned joint actions ("shared competence") between the EU and member states, if required to address common safety concerns in public health matters. Specifically, Article 168 refers to "measures setting high standards of quality and safety for medicinal products and devices for medical use" ["Treaty of Lisbon amending the Treaty on European Union and the Treaty establishing the European Community, signed at Lisbon, 13 December 2007. Article 168, Public Health. http://www.lisbon-treaty.org/wcm/the-lisbon-treaty/treaty-on-the-functioning-of-the-european-union-and-comments/part-3-union-policies-and-internal-actions/title-xiv-public-health.html].So, public health should be added to the list of rights at page 5.Page 7, point 3.5: "citizens should enjoy a right to be informed of any automated treatment of their data by government bodies, and systematically be offered to express opt out. Citizens should never be subject to systematic scoring by government". Actually, GDPR put the attention on the fact that no processing of personal data can be made without specific consent without clearly explaining the kind of processing that will be made of data and its purposes, and this is valid also for government bodies towards citizens. Particular care should be taken to ensure transparency and governance for individual level nationwide, administrative registry data. These data are expected to increase in scope in both degree of detail and magnitude, and can include interactions with the healthcare system, sociodemographic data, potentially genetic data and other sensitive information thereby posing particular risks within the field of trustworthy AI.Transparency, governance and control mechanisms should be established to ensure integrity and respect for the individual when related to use of AI	1. Accountability: in the healthcare domain, clear rules need to be made to define mechanisms to identify responsibility, in case of health problems or missed health benefits due to AI. If the physician trusts the AI system, and this is in error, who is responsible? How the chain of responsibility works? Is it the physician? Or the software house? Or the developer of the AI, or who provided the training datasets? 2. Data Governance: as stated before in the principle of justice, well-balanced (for the aim of the AI) training dataset needs to be used. Transparency on how the training has been carried on, as well as reached performances, needs to be achieved. Integrity of patient-originated anonymous data is in particular critical once used for training: in the data acquisition phase specific means to verify data quality should be put in place if secondary use for AI is forecasted.3. Design for all: a-priori evaluation of the digital (health) literacy of the expected user/patient should be considered into the design.4. Governance of AI Autonomy: when applicable in healthcare topics, human oversight on AI results of paramount importance for reliably interpret the results and correct them when unreliable or unrealistic or unethical. This could apply also to business models for public healthcare systems, or health resources allocation. 5. Testing and Validating: "Testing and validation of the system should thus occur as early as possible and be iterative, ensuring the system behaves as intended throughout its entire life cycle and especially after deployment". It is important that any change that could affect the performance of the AI system is declared to the final user, as well as the reasons behind this change and expected impact on results. As for every medical device, post- market surveillance on AI-based medical devices should be particularly enforced, as well as any failure promptly recorded and made public, to proactively act in modifications or recalls.6. Explanation (XAI research): providing	Page 24: "The primary target audience of this chapter are those individuals or teams responsible for any aspect of the design, development and deployment of any AI-based system that interfaces directly or indirectly with humans, i.e. that will have an impact on decision-making processes of individuals or groups of individuals": Accordingly, we suggest that, in order to prepare the use cases envisioned for the next version, stakeholders specific to their different domains, if not already represented in the high level group, should be involved in the discussion and active preparation process.We agree with the proposed circular model, supposed that all the different stakeholders will be involved, thus avoiding self-referentiality.In addition to the proposed assessment list, in particular for the healthcare scenario, it would be proper to define a risk assessment (and relevant mitigation strategies) in order to evidence which aspects of AI application could introduce higher danger to the patient, so to prioritize higher attention to related ethical problems. Here below a first attempt:AI tools included or to be potentially included into procedures with direct effects on patients:Application Risk1 Research: disease understanding and modeling low2 Diagnosis moderate/high3 Prognosis high4 Therapy selection and patient monitoring high_____ _____In addition, specific information should be made transparent to the potential different users. Here a tentative list:Compulsory disclosure to both Patient and Healthcareprofessional:- Are AI tools present in the medical procedure/practice/device? Is AI tool certified as medical device ? Relevant study results supporting certification should be given, even if unpublished. - What are the scope of the AI algorithm and the use of its output in the context of the whole specific application? - Conformity of algorithms to	Potential of AI in several application fields is still unknown, so there is a real need to anticipate as much as possible the ethical concerns that could arise from its massive use.As regards healthcare applications, ethical concerns are interconnected with privacy, confidentiality, patient-healthcare professional relationship, with possible impact also on medical research.The specificity and the importance of the healthcare sector should lead to the creation of a specific High-Level Expert Group, where representatives of different medical associations and patients are included, in order to properly decline the ethical problems in practical scenarios, and having the health of the patient as only compass to find the right solution.

		Level Expert Group for each domain of AI application, and in particular for healthcare including representatives of different medical associations and patients. Also, experts in theology could be involved, as in ethical committees they happen to be represented to give specific views.	in nationwide, individual level registries. Page 9 – line 2,3 “the principle of beneficence”: “by helping to increase citizen’s mental autonomy, with equal distribution of economic, social and political opportunity”. Healthcare access could be added as a goal for equal distribution.“the principle of non maleficence”: in the healthcare domain, this translates into the need to preserve privacy of the individual, as well as to show potential clinical value connected to the AI utilization larger than the possible patient’s health risks, ,which again is related to transparency of medical devices (AI as software as a medical device).“the principle of autonomy”: in the healthcare domain, this implies careful investigation of psychological mechanisms in the healthcare provider and patient relationship when exposed to utilization of AI, for example for computer-aided decision support system. Page 10 - “the principle of justice”: in the healthcare domain, this can be referred to the need to have well-balanced training sets for the aim of the AI (i.e., same male and female composition, balanced age groups, ethnicities, who generated the gold standard –one center or multiple centers of different countries, etc.), and consequently to be transparent on the reliability of the results specifically related to the training group population or experience of the “gold standard”. For possible problems arising with this, please refer to this article about IBM Watson: https://www.statnews.com/2017/09/05/watson-ibm-cancer/ .“Justice also means that AI systems must provide users with effective redress if harm occurs, or effective remedy if data practices are no longer aligned with human beings’ individual or collective preferences.”: in the healthcare settings, this translates into the need to update and re-train the AI in case medical guidelines are modified. The practice of using synthetic cases to treat and train AI on a variety of patient variables and conditions that might not be present in random patient samples, but are important to treatment recommendations, should be discouraged until a proof of equal reliability, accuracy and clinical value using real cases is produced [see again arising problems with IBM Watson: https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/].“the principle of explicability”: in the healthcare domain like computer-aided decision support systems, besides technological and business model transparency, evidence of clinical value has to be given, which means complete transparency on the testing protocols, dataset composition and origin, validation results, and other possible information that could help in understanding the benefit-risk ratio of using AI as predictive tool, in particular in critical scenarios (which would imply class IIb certification of AI software as medical device). It appears of paramount importance the transparency of AI processes that led to a specific diagnosis or suggested treatment, not limited to the link to a specific literature, but why that choice was recommended for that particular patient.Page 11 - 5.1 Identification without consent: in the healthcare domain, patient confidentiality and privacy are at the basis of the trusting relationship between the patient and the healthcare professionals. It is of	explanation about system behavior in a simple way is the necessary step to create trustworthiness for introduction of AI in the healthcare domain, in particular where treatments are suggested or patients are classified according to a pathology or a risk score.In the healthcare scenario, additional non-technical methods that can be considered in order to address the requirements of Trustworthy AI should include the support (and requirement) to perform extensive longitudinal and multi-centric randomized clinical studies to prove its accuracy and clinical value, in particular for unsupervised AI methods where a gold standard is not available.	the ethical principles and the requirements of the Trustworthy AI as in the final guidelines - Conformity of data management to the EU GDPR Compulsory disclosure to the Healthcare professional:- Detailed risk analysis- What is the AI method used?- What is the nature of data used for training /test? Number of datasets, geographical origin, details about - What was the strategy to obtain an unbiased training dataset? - Which criteria were used for the evaluation of the algorithm performance and final performance score? - What are the parameters/features extracted from the raw input data actually considered by the AI algorithm?

				paramount importance to preserve this, in order to prevent abuses that could generate disparity of treatment (i.e., not employing a person due to his/her health background check, or determining higher prices for insurance) or even discrimination (i.e., HIV positive patients).If “anonymous” personal data can be personalized, then their treatment and processing falls automatically under the current GDPR that implies explicit consent to process that data.A recent paper [Liangyuan Na et al. Feasibility of Reidentifying Individuals in Large National Physical Activity Data Sets From Which Protected Health Information Has Been Removed With Use of Machine Learning, JAMA Network Open (2018). DOI: 10.1001/jamanetworkopen.2018.6040] highlighted this problem, showing that it was possible to reidentify 20-minute-level physical activity data from 14 451 individuals that have had protected health information removed by using machine learning with accuracy of approximately 80% in children and 95% in adults, thus highlighting the need for better protection and related data misuse [https://techxplore.com/news/2018-12-advancement-artificial-intelligence-health-privacy.html].5.2 Covert AI systems: in the healthcare domain, in particular in telemedicine/teleconsultation settings, it will be of paramount importance to clearly understand if interacting with a real physician/nurse or with a machine.Page 12 - 5.3 Normative & Mass Citizen Scoring without consent: as for identification, mass citizen scoring should not be applied for healthcare purposes in a limited social domain, as it potentially labels single individuals based on their possible pathology, thus potentially breaking privacy and creating disparity.5.4 Lethal Autonomous Weapon Systems: as the benefit/risk ratio in case of error is the death of an innocent person, this utilization should be banned forever, also in line with the human-centric principle claimed at the beginning of the document. 5.5 Potential longer-term concerns: a potential concern could arise from the massive utilization of such systems in the healthcare domain, thus preventing possible discoveries of new diseases (the AI system would not have been trained for it), or limiting capability of decision of medical professionals if they just have to trust the AI. Accountability and traceability, as well as long-term randomized clinical studies would be paramount to observe and thus act in preventing such behaviors.	
Patrik	FLOREEN	University of Helsinki			I already gave comments, but would like to make a further clarification concerning my text about explainability. My text was: “Explicability could be mentioned explicitly. Now there is a mention in passing about users “having the facility to interrogate algorithmic decisions in order to fully understand their …” I disagree about the “fully understand”. People do not ask how google work when they google, and they need not know the technical details about what is happening. The same for their car navigation system. Do they need to know how GPS works and how the AI calculates the fastest route? They don’t. They only need a vague understanding of what is going on.”

						The point I want to make is that there are different types of explanation. The users do not need a technical explanation of what is going on; this is what I was referring to. So they do not need to know that Dijkstra's algorithm was used to calculate the optimal route from A to B. They do need, for instance in case of automatic decision making, an explanation about what factors were relevant and how for the decision, in particular, of course, if there is a legal requirement to provide such information. The technical details are needed only for experts to be able to verify that the system works as it should. Thus, that is a question of auditability. The systems should be auditable.	
ALBERTO	DIEZ-OLIVAN	TECNALIA R&I	In EXECUTIVE SUMMARY, page i, it's mentioned that 'To ensure that we stay on the right track, a human-centric approach to AI is needed, forcing us to keep in mind that the development and use of AI should not be seen as a means in itself, but as having the goal to increase human well-being.', whereas in the same section, in 'A. RATIONALE AND FORESIGHT OF THE GUIDELINES' in page 1 ('Trustworthy AI'), it is said that 'AI is thus not an end in itself, but rather a means to increase individual and societal well-being.'; which is a bit confusing to me. Maybe using different words would help to clarify whether AI is a means or a goal, as a paradigm by itself.	In footnote 2, page 6, one of the following words should be removed: 'Our' or 'the', in sentence: 'Our the use of values here'. In the bullet 'The Principle of Non maleficence: "Do no Harm"', page 9, it is said that 'Avoiding harm may also be viewed in terms of harm to the environment and animals, thus the development of environmentally friendly'; in this context, the key concept 'animal welfare' is of special interest when talking about animal farming and AI. Maybe this concept should be mentioned there, besides the 'environmental awareness', due to its importance. In the bullet 'The Principle of Justice: "Be Fair"', page 10, would it be possible and feasible to identify or even further define a legal entity or similar that will be on charge of ensuring that such Principle is actually respected? This also refers to the 'Critical concerns raised by AI', from page 11 onwards: indeed, similarly as the GDPR in the case of data protection, an official regulation and/or legal entity is needed regarding AI. With relation to LAWS concern, it's a tricky issue, but the legal responsibility of actions should be on the person/s that decide to use the corresponding AI system in a harmful way (i.e. governments, armies or military groups); anyhow the design of the AI system must be driven by the idea of not harming anyone. It's like when discussing about guns and more traditional weapons: the person who pull the trigger is the responsible of the resulting action. The last point, 5.5 Potential longer-term concerns (page 12), is also complex, but to me all AI systems should evolve and self-learn within certain logical limits and always under human supervision.	In 'Data Governance' (page 14), some important concepts regarding this issue are not mentioned, such as overfitting or online learning (for self-learning systems, for instance). Another important decision to make is when and how to retrain a model given new unseen data and events (model drift). It must be noted as well that once a model has been properly trained, tested and validated, such model could be reused in a similar scenario even if there are no data available yet (domain adaptation or transfer learning), in order to provide outcomes from the very beginning but with less accuracy. Footnote 24 in page 15 is quite interesting. Maybe it could be included in the text. In Subsection 8, 'Robustness' (page 17), the paragraph related to 'Fall back plan' could include something about version control systems (i.e. Git and SVN) to ensure that a functional, optimal version of the model (the serialization of the corresponding object) is always available. In 'Testing & Validating' bullet (page 19), within the last paragraph the so-called 'acting-module' is introduced; the definition of real, automatic actuators driven by AI systems outputs must be very carefully considered, for instance when dealing with assets (i.e. heavy machinery) that may threaten the safety of workers and humans in general depending of which actions are taken. In 'Education and awareness to foster an ethical mind-set' I would emphasize the idea of providing AI-related education at school, since children are the future and they must understand what are the main advantages, risks and drawbacks of a society enhanced by AI systems and technology. Actually, I'm personally aware that in many schools there are subjects focused on AI. It clearly becomes a cultural and educational issue for the acceptance of AI in the near future.	It is very difficult to define generic assessment questions that are applied to any use case and sector. I think some questions could be domain-specific, like open questions to be defined by the concrete use case actors and evaluated and validated by a legal entity and/or regulator. A use case focused on Industry 4.0 is missing.	In general terms the document is very readable and easy to understand, at least for readers with some knowledge on the AI paradigm. The key points are well explained and discussed. The 'key guidance boxes' at the end of every section are very useful to reinforce the ideas and concepts that are transmitted.

Mauritz	Kop	Artificiële Intelligentie & Recht	none		Page 10, Principle of explicability: I think this is currently a very hot topic. Transparency is often lacking in systems that automatically generate decisions, as is the case with certain authoritative or administrative bodies. Algorithms are quite frequently used to generate decisions in administrative matters. It has become increasingly unclear as to WHY those algorithms come to a certain outcomes, what patterns underly the outcome and HOW those patterns cane about. There is also often uncertainty about the procedural guarantees because of this.Page 11, Identification without consent: very important topic in the context of 'harvesting' training data. There are stakeholders that are absolutely sure that the governmental bodies should harvest as much data from citizens as possible, in order to train AI. E.G. collecting data from wearables and smart apps. Needless to say this poses great problems in terms of privacy but also in terms of autonomy.Page 12, Normative and Mass Citizen scoring without consent in deviation of fundamental rights. @ “However, whenever citizen scoring is applied in a……of the scoring mechanism”: Sure, but a very important question here is how you aim to assure that such consent is really and truly informed. As pointed out by the HLEG on page 11, the current mechanism is to give consent without consideration. Providing the 'average' citizen with heaps of information about processes, purposes and methodology is not going to change this.Page 12, @ 5.4 Lethal Autonomous Weapon Systems: Very unfortunate choice in the order of words! I’d suggest they be swapped around. Autonomous Lethal Weapon Systems (ALWS) would be much better, especially since you then create a clear difference between LAW and UNLAWFUL.Page 13, @ footnote 18: I have very strong sympathies for Searle's theories here (Chinese Room Experiment). I think it is ultimately IMPOSSIBLE to attribute consciousness to machines- however clever they are- because they simply lack the physical processes a brain goes through in order to create what we see as consciousness. Even IF one succeeds in creating a machine that strongly resembles conscious processes, this does NOT mean that the machine is conscious or could ever be conscious.	Page 21, standardization: co-regulation by an independent (governmental?) body! Accreditation systems should be independent. I think we should avoid appointing (private/commercial) notified bodies here.Page 22: For both accountability and codes of conduct: all very applaudable, great tools in order to create awareness, but there should be some kind of mandatory enforcement mechanism. Self regulation is - without a doubt- a very important tool, but in the case of AI a bit more is needed, especially because AI has so many (potential) implications for human rights.Page 22, @ stakeholders and social dialogue: The benefits of AI COULD be many, provided that we do not end up in endless discussions between stakeholders. Connecting various participants should be key here.Page 22, @ diversity and inclusive design teams: I would very much like to know if the HLEG has ideas about the practical execution.	The final version of the EU AI Ethics Guidelines for Trustworthy AI can be incorporated in the AI Impact Assessment.https://ec.europa.eu/futurium/en/european-ai-alliance/ai-impact-assessment-available-english-code-conduct-and-roadmap-included	This is mere a thought, but it has crossed my mind regularly over the past few weeks. Just assuming that a level of consciousness can be attributed to AI, then I think it would be necessary to set up a legal framework that resembles the current GMO-regulation, in order to regulate the actions with AI . Another thing is that we should be careful not to overestimate the possible/future implications of AI for the average citizen. We thank you for your important work and for the opportunity to provide feedback!Kind regards,Mauritz KopSuzan Slijpen

Stefan	Keller	Private	Realistically future solutions will consist of several integrated AI components, that might have different pedigrees. The corresponding data processing activities will not make use of one Trustworthy AI, but rather an integrated solution of such components. I'm missing this aspect. The Introduction fails to differentiate the main stakeholders: a) The manufacturers of AI components b) The manufacturer (integrator) of the solution c) The data controller making use of such a solution. GDPR puts the regulatory burden on the data controller, who in most cases is not able to get sufficient guarantees on the details and privacy compliance of the solution. This is especially true in situations in which the solution is provided to individuals as data controllers, e.g. in the form of a software or integrated device. These data controllers can include children and elderly. Data controllers will typically not have "golden records" at their disposal to check AI components for bias. Example: Companies buying a HR solution, that comes with integrated application ratings, will have to rely on the Software vendor for the Trustworthiness of the AI (incl. accuracy and non-bias). In my personal view, Trustworthy integrated AI solutions should come with a) a list and pedigree of all AI components b) some certification for each AI component c) privacy seal for the integrated AI solution I believe that the manufacturing certain types of AI components and the integration of certain Trustworthy Integrated AI solutions should require a license or be covered by an extension of the privacy-by-design principles in GDPR. (Today, manufacturers - who are neither controller nor processors - are only indirectly covered.) ---- Some more detailled comments: page 3 : AI recommending medical treatment would most likely be a Software-as-a-Medical-Device (SaMD), and covered under separate existing regulations. These could be a starting point for the discussion.	The discussion in this chapter is overruled by the Data Protection Impact Assessment approach mandated by GDPR, which focuses on impact to the rights and freedoms of the affected individuals - by data processing activities. - Data processing activities might make different uses of AI components, very much the same way a hammer can be used to build a house or kill someone. I believe your discussion here falls short of what's actually expected on the basis of GDPR. A trustworthy hammer in itself is not "ethical" or "unethical". Subsection 4 - in my mind - should really be for the data processing activities (not the AI component or solution). Subsection 5.1 has been covered at length by GDPR, prior A29WP opinion and current EDPB guidance. (Data Protection Impact Assessments, as well restrictions on profiling, ..). - However, my comment on "Introduction" applies, as many data controller will literally "not know what they are doing" with the products in their hands. (--> need for certifications and seals, license to manufacture certain products) Subsection 5.2 - Use of covert AI would violate the "reasonable expectations" of data subjects. However, I doubt that - except for processing based on legitimate interests - this is currently adequately covered in GDPR. In my view, any AI components and solutions must identify themselves. Subsection 5.3 - The household use of (standalone) AI products by individuals is an ugly twist here. GDPR does not apply, neither to individual user nor the product vendor. Potentially massive impact on affected individuals, e.g. by detecting "gays" or "jews" based on some AI in watches or smartphones. (Local software without cloud, etc..) This might need regulation. Subsection 5.4. - also true for cyber warfare. Subsection 5.5. - based on my personal and professional experience it is not true that all current AI is still domain-specific or that it would need scientists. AI components are available on the market for integration - incl. e.g. recognition of unknown faces in home CCTV, voice recognition, job applicant screening, etc. This subsection should include the unknown risks of combining AI components from different pedigrees, either for AI training or product integration reasons.	I believe this chapter should focus on data processing activities making use of AI - rather than focusing on an AI component in isolation. Section 1: Subsection 1: It is completely unclear if you if and what accountability you see on developer, product manufacturer, data controller or end user. In the discrimination example, moral, physical and material harms might exist that are clearly not remediated with an explanation or apology! (See e.g. CNIL PIA guidance) Subsection 2 - data quality is very hard in AI research. It is unclear which of the stakeholders have which responsibilities here. You do not address the risk of learnt personal data being extracted by an adversary. Subsection 3 - I strongly disagree that all AI-based solutions should be available to all. For example, some clinical solutions must be reserved to trained Health Care Personnel. Other solutions must clearly be reserved to Law Enforcement and National Security. - These points are already covered in existing laws. Subsection 5 - This is a requirement on the data processing activity - NOT on the AI. For example, an AI can be used to diagnose a disease status - which can lead to discrimination. Subsection 6 - Very valid and important point. Not covered via GDPR today. Very relevant. Subsection 7 - Please specify how that applies to the different stakeholders! An end-user with an AI-powered discriminating product in his hands will - in the absence of a processor, and under the household exemption - not fall under GDPR! (Manufacturer is not covered by GDPR). "stages of the life cycle of the AI system" is not defined. - Development, commercialization, retirement? or a Plan-Do-Check-Act cycle on the end-user side? Section 2: A Lifecycle for an AI component or A Lifecycle for an AI-based solution? This discussion falls short of GDPR expectations on Data Protection Impact Assessments. While I welcome the development of certification criteria for AI and the categorization of AI to set restrictions on its development - I feel this section falls short on what it intends to achieve. Who is the intended audience here? - as it feels like a self-discussion between researchers, without regards to the views of developers, integrators, controllers or end users. The idea that regulatory bodies would undertake verification and auditing of AI systems seems far-fetched. - AI-based solutions are already present in Software-as-a-Medical-Device. It should be possible to map these discussions to this already well-regulated field. Subsection 2 - are you aiming at the analogue of pharmacovigilance for AI? - This is a good idea, but would entail traceability of the different AI components and their pedigree. Example: AI component	This is useless. There are several lifecycles: 1. for the AI component 2. for the actual (integrated) solution (product or service) 3. for the actual deployment and operation by user or controller So you need several checklists - and need to clarify the different roles! As part of my day job, I often review Data Protection Impact Assessments on solutions that integrate AI. Many elements of this list wouldn't work out in practice. The original AI developer is in many cases not the data controller nor manufacturer. Also - "6. Respect for Privacy: If applicable, is the system GDPR compliant?" - Data processing activities can be GDPR compliant - whereas products (system) can be designed in a way that allow their operation in a GDPR compliant way. So for an isolated AI system, this question will in most cases not make sense.	I'm very disappointed by this paper, as it does not address my key concerns with AI-based solutions. Realistically future solutions will consist of several integrated AI components, that might have different pedigrees. The corresponding data processing activities will not make use of one Trustworthy AI, but rather an integrated solution of such components. I'm missing this aspect. The Introduction fails to differentiate the main stakeholders: a) The manufacturers of AI components b) The manufacturer (integrator) of the solution c) The data controller making use of such a solution. GDPR puts the regulatory burden on the data controller, who in most cases is not able to get sufficient guarantees on the details and privacy compliance of the solution. This is especially true in situations in which the solution is provided to individuals as data controllers, e.g. in the form of a software or integrated device. These data controllers can include children and elderly. If these users operate AI-based devices under household exemption, GDPR will not apply. Data controllers will typically not have "golden records" at their disposal to check AI components for bias. Example: Companies buying a HR solution, that comes with integrated application ratings, will have to rely on the Software vendor for the Trustworthiness of the AI (incl. accuracy and non-bias). In my personal view, Trustworthy integrated AI solutions should come with a) a list and pedigree of all AI components b) some certification for each AI component c) privacy seal for the integrated AI solution I believe that the manufacturing certain types of AI components and the integration of certain Trustworthy Integrated AI solutions should require a license or be covered by an extension of the privacy-by-design principles in GDPR. (Today, manufacturers - who are neither controller nor processors - are only indirectly covered.) Why didn't you reuse existing EMA guidance on Software-as-a-Medical-Device? Why didn't you leverage the Data Protection Impact Assessment in GDPR? Why didn't you clearly distinguish the various stakeholders, their area of influence and responsibilities and lifecycles? -- I do agree on the need for AI to declare itself as AI to humans.

					FaceDetectA is flawed, so fix in Product A, B, C, D. Standardization would need to extend to non-EU component vendors, as e.g. there is massive AI component development in China. Need for seals similar to CE mark. Accountability Governance must also be with product manufacturer. GDPR does not apply in absence of controller/processor, if under household exemption - e.g. individual end users with a device.		
Isabelle	Schlegel		Section: Trustworthy AI Sentence: "Indeed, even with good intentions or purpose, the lack of technological mastery can cause unintentional harm." Comment: But unintentional harm, effects and misuse of technologies can always happen. Not only due to the lack of technological mastery. From a media perspective, each medium - which compromises all integrated, implemented and applied technologies - co-determines its use. The same applies to the sentence: "Moreover, they should take precautions that the systems are as robust as possible from a technical point of view, to ensure that – even if the ethical purpose is respected – AI does not cause unintentional harm." Comment: Unintentional harm can always happen.	Section: Critical concerns raised by AI Comment: This section should also treat the point "un-intentional harm" and should deal with the topic around technology/ media mis-use. Again, from a media perspective, each medium - which compromises all integrated, implemented and applied technologies, medium equal use of technology - co-determines its use. It is very likely that new technologies and AI systems are used in a way which was not intended. That the system is applied and embedded into social practice where people will use the technology differently than planned and un-intended effects would happen. For example, the robot seal "Paro" which was introduced in elderly care can be "un-intendedly" used when the caregiver goes out for a five minute smoking break. Then the patient is left alone with a machine. This may not be critical, but imagining a further development of human - machine - interaction, we need to answer the question how we want to be treated. Another example: The hope for more democracy through platforms like Facebook showed its downside when the US election was influenced via those platforms and channels.			
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential

Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Stephane	Senecal	Orange		§4 Ethical Principles in the Context of AI and Correlating Values The Principle of Non maleficience: "Do no Harm" (page 9) End of first paragraph: "AI systems should be developed and implemented … algorithmic determinism": please explain what is meant exactly by "algorithmic determinism". The Principle of Explicability: "Operate transparently" (page 10) First paragraph: "AI systems be auditable, comprehensible and intelligible by human beings…": this requirement seems very difficult to achieve for state-of-the-art deep learning based systems which are currently deployed in products and services.	§1 Requirements of Trustworthy AI 8. Robustness and/or 9. Safety (pages 17 and/or 18) I would suggest to add/mention the exploration/exploitation trade-off or dilemma in interactive (reinforcement) learning based systems. This issue can lead to serious and unavoidable system performance degradation during the learning phase and should be properly addressed within the design of algorithmic learning methodologies and solutions, before deployment.		
GEORGIOS	LEKKAS	Representation Office of the Church of Greece to the E.U.		Archpriest Dr Georgios LEKKAS Counsellor at the Representation Office of the Church of Greece to the E.U. (Brussels) ARTIFICIAL INTELLIGENCE AND AN APPLIED ETHICS. BUT WHAT KIND OF ETHICS? We warmly welcome the initiative of the European Commission and the High-Level Expert Group on Artificial Intelligence to hold a public debate around the Draft Ethics Guidelines for Trustworthy AI which should be observed by designers of artificial intelligence systems for a 'credible artificial intelligence made in Europe'. The ever more advanced systems of artificial intelligence, which inevitably promote an implicit ethical view, will through their repeated use become a means whereby generations of European citizens are educated. For this reason this is indeed a commendable attempt by the European High-Level Expert Group on Artificial Intelligence to set out in a Statement of Principles the ethical values and principles that the designers of primarily European artificial intelligence systems should establish and promote. The authors of this draft have rightly undertaken to derive a system of ethical principles and values from the current legal framework on fundamental human rights, as stated in the EU Treaties and the European Charter of Fundamental Rights, and then to advocate its implementation by everyone involved in the operation of artificial intelligence systems. However, the European legal system on human rights can be interpreted either within the framework of an individualistic morality, designed to protect the individual from the society in which he lives, or through a collective ethics by which fundamental rights are recognized in every human being as a necessary prerequisite that permits him to live in a society of peace		

and love with everyone else (= the social principle).

The proposed European draft ethics for artificial intelligence presupposes the conception of the human being as an autonomous rational and free entity, who is obliged to engage with his or her counterparts only in order to serve his or her own complex social needs. However, this proposed draft of the ethics, based on the philosophical model of a human as the self-referential being par excellence, being autonomous and ontologically sufficient in itself, extends and applies to the field of artificial intelligence - and before long also autonomous artificial intelligence - the individualistic conception of the human being whose predominance on a global level has already produced devastating effects for every individual human being who faces the spectre of isolation from others as well as the destruction of our planet.

Is the idea of the 'autonomous' human being to which the draft under discussion often refers adequate for an 'human-centric' approach to artificial intelligence? As an Orthodox Christian my answer is 'no'. The experience of two thousand years of Orthodox tradition says that human beings are not merely autonomous rational entities who relate to others out of the need to survive, but free and intelligent loving hearts which by virtue of their own nature require others in order to be free. From this perspective, the others (our Creator-God, fellow human beings, the Cosmos) are necessary for my freedom, simply because without any of them I will not have any choice at all. In Orthodox tradition, human beings are not considered to be units that need to coexist peacefully within society merely because this serves the separate individuality of each of them; instead they are conceived of as members of a common body where the condition of each member necessarily affects the health of the whole body and the health of the whole body has beneficial consequences for the proper functioning of each member - that is what I mean by the term "social principle".

For Orthodox tradition, human beings are free to think or not, they are free to love or not, to act or not, but their freedom cannot be formulated in algorithmic terms because it exists prior to reflective thought, since at heart it is a consequence of the ex nihilo creation of mankind, that is, the origin of mankind solely and exclusively in the free creative will of our Creator-God. That is why the human being is not in danger of being destroyed by autonomous systems of artificial intelligence, from which human intelligence seeks to protect itself through Guidelines such as the one we are now discussing. Since the mystery of human freedom - but also of the dynamic entity that constitutes the human being - is hidden in our deep and ontological relationship with others, and especially with our Creator-God, human beings are only in danger of being destroyed by themselves. For the possibility of such an outcome being ruled out today, it is not enough to draw up ethical guidelines for artificial intelligence; we must, among other things, immediately criminalize investigations into the merging of human capabilities and machines as crimes against

humanity which must be punished by the gravest of the penalties provided for in our European legal system. We consider it equally urgent that legislative initiatives are ratified within the EU to prevent the anthropomorphic simulation of artificial intelligence systems to such an extent that it becomes difficult to discriminate between - or even establishes a societal belief in the equivalence of - human and machine.

The question is not, however, how to prepare ourselves to resist the impending autonomous systems of artificial intelligence, but how to use them in the service of our ontological interrelation with everyone and with everything (our fellow humans, the Cosmos, our Creator-God) in order to attain a bliss from which the machine is excluded, by virtue of its nature. We are human beings - anthropos is the Greek word from which the English words anthropology and anthropological are derived - because we have been made to live in relation with others, first with our Creator and God – "ano" as a prefix of the Greek word anthropos means someone or something which is higher than we are - and then with all the other human beings, brothers and sisters, regardless of colour, race or religion. Therefore, in accordance with such an understanding of the human as an essentially relational being, we have been created to love in freedom and with all our heart and with all our mind our Creator-God, as well as to cherish in freedom all other human beings, just as we should love ourselves - the two prerequisites of our bliss.

If human happiness presupposes a deep and lasting association with others, the recognition of the 'social principle' as a necessary principle for the operation of artificial intelligence systems is essential. Such a principle dictates that the operation of these systems ALWAYS serves, in the short or long term, the ontological need for a deep coexistence between all human beings within the single body of mankind, otherwise any such systems will be rejected, since each time the relational value of the human being is put in danger, the human being risks, willingly or not, being turned into something far inferior to man, into a beast or a man-machine. The technical and non-technical methods which need to be called upon for the application of the social principle for the design and operation of artificial intelligence are the task of the scientific community. Nevertheless, the question of what sort of ethics is applied in the area of Artificial Intelligence must be a decision arrived at through the broad consensus of civil society and its organizations and the cooperation of intellectuals and scientists – for this reason we warmly welcome this debate as a necessary step in the right direction - so as to avert the risk that the individualism which Europe has inherited from the previous century is placed on a pedestal, and results in a new situation in Europe where my other half is my robot!
Brussels, 21.1. 2019

Sofia	Stigmar	Swedish Trade Federation	The guidelines could be used as a certification or signal to consumers that a company complies with the ethical guidelines. In that case it is of outmost importance that the guidelines are clear and concrete to ensure correct implementation by the companies. When guidelines like these are presented, companies tend to see them as mandatory and it is therefore important that the document is instructive and accessible. One additional concern is whether the guidelines should be viewed as a competitiveness tool? In other words, is compliance with these ethical guidelines intended to be considered a competitive advantage. If that is the idea, it is even more important that the guidelines clearly state what must be done to become compliant to guarantee fair competition. To sum up, The Swedish Trade Federation want to clarify that our analyze rather shows that the guidelines are in line with what the customers demand, and that acting according to the guidelines is something positive for our member companies to be competitive.	Within the EU, we should always start from the set values and ethical codes that apply throughout as a foundation. Therefore, AI must also be set up with its values as a warrant and "defalult". For example, AI may never be used for discrimination, harassment and personal privacy must always be respected and AI must never be used so that it becomes conflicts of interest or corruption. A common understanding of fundamental rights is key to ensure a harmonized implementation of these guidelines. Different interpretations may result in different solution, especially when two fundamental rights collide, which will happen while developing and using AI. For example, there sometimes need to be some harm to develop a new AI system. Therefore, we call for an addition in the document with guidance on the hierarchy. In short, which rights and interests have the priority? We would also like to highlight that there is a confusion in the use of the words “user” and “consumer”. The companies are many time the users since they have installed the AI technology in their sales-processes. The users in the present wording are rather the individuals who are for example buying goods on a website, in other words the consumers. Critical concerns raised by AIThe Swedish Trade Federation finds the heading of this paragraph problematic. It is obvious that this chapter aims to identify actions that are not compliant with ethical behavior in AI, even though it is just called critical concerns. When a high-level expert group, initiated by the European Commission, points out several actions that are “critical concerns”, many SMEs who lacks the economy, knowledge and other resources to investigating the issue themselves, will follow the groups guidance. It is therefore important that the information is correct and comprehensive. For example, the Swedish Trade Federation does not agree that it is always forbidden to use AI for identification without consent. Neither do we agree with the statements on converted AI systems. Today, many of the AI systems are embedded in other technology which makes this statement much to simplified. However, the Swedish Trade Federation contends that whether the consumer detects that he or she is talking to an AI system or a human is something that will not be as crucial in the future, why this should not be stated as a mandatory service in ethical guidelines that are supposed to be technology neutral.	The basic idea of the chapter is good, but it must be more concrete. The design of the chapter in its present form is mostly focused on problematizing instead of giving guidance. We encourage the parts that aim at increasing transparency as it is something we are convinced will increasing the confidence of the system for the consumer. It should be clear where in the sales chain AI is used and how. As a result, AI becomes a natural feature of the lifecycle	The introduction of an assessment list is very good and helpful for most of the actors on the market. We call for a clarification that underlines that all 10 points in the assessment list cannot and does not have to be operated by all companies because of different requisites. Although the general approach is good, we do identify some room for improvement:• There is sometimes more than one actor responsible when things go wrong with AI. This needs to be emphasized in the chapter on accountability. Overall, we ask for a deeper account of the lifecycle. Many times, the AI user is neither the developer nor the ultimate user. • There is also a risk of pointing out one accountable if something goes wrong if what went wrong is something that the part could not be hold accountable for. The principle of legal security is very important. And we call for a deeper understanding of the full lifecycle.• Not all companies have the resources to establish a review board of ethical AI. We understand that the document has a broad perspective – but we ask for a clarification that underlines that all these points in the assessment list cannot and does not have to be implemented by all companies because of different prerequisites. • If the system must be designed to accommodate all special needs or disabilities it could be very costly to the provider, and sometimes even impossible. • The non-discrimination points should not be phrased so they hinder a company from contractual and industrial freedom. For example, customer segmentation is something many companies do today with AI. Will this not be compliant with the guidelines? • On the matter of information, it cannot only be the responsibility of the companies that are using the AI-systems to describe technical parts and how the processes are taking place. Companies must be able to rely on a certain amount of prior knowledge by the consumer. This knowledge needs to be acquired in school or through the public information campaigns, otherwise the information burden will be too heavy on companies. Another aspect of this is also that sometimes it is impossible for the user to describe the technology behind the system since they are not the ones who have developed the AI. • There is always a risk when using a wording like “clearly communicated” - what does that mean?	First, the Swedish Trade Federation welcomes the idea of AI Ethics Guidelines and agree that there is a need for further guidance in this new area of technology. We strongly believe that AI is key when addressing many of the major challenges that businesses are facing now and in the future. We are also convinced that it is important that the guidelines presented and produced by the European Commission’s High-level Expert Group on Artificial Intelligence strive to maximize the benefits of AI while at the same time minimizing the risks. It is important to underline that entrepreneurs and not the legislators are the source of new technology.Furthermore, we already see negative developments in countries around the world where AI is used without safeguarding fundamental rights. This is not a desirable development, and it is crucial that the EU leads by example and show how development and fundamental rights can be combined.Moreover, the Swedish Trade Federation question whom this document is directed to? In its present form it is more educational than practical. If the document is to be useful for companies, we think that there must be limited discretion.

Nathalie	NEVEJANS	University of Artois (France)	> Page 2: We would like to stress that the following phrase seems either inaccurate: “The goal of AI ethics is to identify how AI can advance or raise concerns to the good life of individuals, whether this be in terms of quality of life, mental autonomy or freedom to live in a democratic society. It concerns itself with issues of diversity and inclusion (with regards to training data and the ends to which AI serves) as well as issues of distributive justice (who will benefit from AI and who will not).” One definition of ethics by Merriam Webster is: “the discipline dealing with what is good, bad, with moral duty and obligation; a set of moral principles; a theory or system of moral values” (Merriam Webster, https://www.merriam-webster.com/dictionary/ethic). The question at hand is not to define what is the “good life of individuals” but rather set the line between good and bad which is a far larger objective than what the HLEG expresses and includes individuals with “bad” quality of life such as the mentally and physically challenged citizens. We believe ethics should not give a clear-cut decision on what a “good life” is. As reminded by the Oviedo convention, fundamental rights are the foundation to ensure the “primacy of the human being” in a context of technological evolution. We propose that this sentence read as follows: “The goal of AI ethics is to identify how AI can advance or raise concerns to the primacy of human beings over technology, to ensure respect for human rights such as freedom of being who they are by virtue of being humans. This leads to the ethical principle of autonomy which prescribes that individuals are free to make choices about their own lives, be it about their physical, emotional or mental wellbeing (i.e. since humans are valuable, they should be free to make choices about their own lives). In turn, informed consent is a value needed to operationalize the principle of autonomy in practice. Informed consent requires that individuals are given autonomy to live in a democratic society. It concerns itself with issues of diversity and inclusion (with regards to training data and the ends to which AI serves) as well as issues of distributive justice (who will benefit from AI and who will not).” In Purpose and Target Audience of the Guidelines : > page 2 "A mechanism will be put in place that enables all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis. This will be set out in the final version of the document". This point is important. Precision will be necessary to specify the obligatory value of the Guidelines. If it is not obligatory, there is a risk that companies may find that taking ethics into account will undermine innovation and profitability. It will also be necessary to provide a means to value the companies that approve the Guidelines.	In 3. Fundamental Rights of Human Beings : > page 7 “At the same time, citizens should enjoy a right to be informed of any automated treatment of their data by government bodies, and systematically be offered to express opt out” How do citizens access services when they opt out? How they still be served by AI based companies when they decide to refuse the algorithm-based decision-making process? In comparison, user experience after the adoption of GDPR, certain websites totally block their access to their content upon refusing cookies. Even if technical cookies are accepted by the user, some companies still refuse to serve the user with its goods and services. We shall see the same bottle neck effect when users reluctant to AI-based decision will not be able to perform their commercial activities as well as their administrative tasks. the expulsion of some users will be inevitable and will de facto undermine the notion of freedom of choice and autonomy. In 4. Ethical Principles in the Context of AI and Correlating Values : > page 8 “It should also be noted that, in particular situations, tensions may arise between the principles when considered from the point of view of an individual compared with the point of view of society, and vice versa. There is no set way to deal with such trade-offs. In such contexts, it may however help to return to the principles and overarching values and rights protected by the EU Treaties and Charter. Given the potential of unknown and unintended consequences of AI, the presence of an internal and external (ethical) expert is advised to accompany the design, development and deployment of AI. Such expert could also raise further awareness of the unique ethical issues that may arise in the coming years.” This paragraph is unclear: which experts? Reporting to whom? Could you please clarify on what unlikely tensions you refer to? In 5.3 Normative & Mass Citizen Scoring without consent in deviation of Fundamental Rights : > Page 12 “We value the freedom and autonomy of all citizens. Normative citizen scoring (e.g., general assessment of “moral personality” or “ethical integrity”) in all aspects and on a large scale by public authorities endangers these values, especially when used not i situations, tensions may arise between the principles when considered from the point of view of an individual compared with the point of view of society, and vice versa. There is no set way to deal with such trade-offs. In such contexts, it may however help to return to the principles and overarching values and rights protected by n accordance with fundamental rights, or when used disproportionately and without a delineated and communicated legitimate purpose. […] However, whenever citizen scoring is applied in a limited social domain, a fully transparent procedure should be available to citizens, providing them with information on the process, purpose and methodology of the scoring, and ideally providing them with the possibility to optout of the scoring mechanism.” First, why a “limited social domain”?	In 1. Requirements of Trustworthy AI 1. Accountability : > Page 14 “Good AI governance should include accountability mechanisms, which could be very diverse in choice depending on the goals. Mechanisms can range from monetary compensation (no-fault insurance) to fault finding, to reconciliation without monetary compensations. The choice of accountability mechanisms may also depend on the nature and weight of the activity, as well as the level of autonomy at play. An instance in which a system misreads a medicine claim and wrongly decides not to reimburse may be compensated for with money. In a case of discrimination, however, an explanation and apology might be at least as important”. This is a very important point. The "accountability" must be distinguished from “liability” (ie. to be legally responsible). Point 1 should establish the distinction between accountability and liability. In . Requirements of Trustworthy AI 4. Governance of AI Autonomy (Human oversight) : > Page 15 We have two remarks : “The level of autonomy results from the use case and the degree of sophistication needed for a task”. This is agreed upon. We recommend an idea by Researcher Cyrille Dalmont (in Intelligence artificielle et santé : 10 propositions anti -brouillard pour régulation éclairée). The user should be informed on the level of AI used to reach a decision by allowing for the identification of diagnoses and prognosis made by artificial intelligences. For this purpose, a pictogram could be affixed on any document, image or prescription produced by an AI. The patient would then be able to identify the degree of human involvement in the conclusions made to the medical examinations carried out and have a recourse if need be. Similarly, autonomy of decision lies in the anonymity of data. Cyrille Dalmont proposes : “The collection and processing of patient data is a crucial issue. risk could simply be color-coded based on the degree of confidentiality or sensitivity of the data and their treatment with state-level labeling of companies and their level of entitlement to process certain data according to precise specifications and security guarantees provided by authorized public companies and organizations. By way of illustration, the data enabling predictive medicine to be carried out should be classified as the most sensitive with an absolute ban on dissemination to certain institutions or economic actors such as insurance companies, banks or lessors in order to avoid Digital precariousness. An individual could no longer get access to insurance, medical treatment, contract a loan or rent a home if his/her risk factors were too important.» Second remark : "FOOTNOTE 24. AI systems often operate with some degree of autonomy, typically classified into 5 levels: (1) Domain model is implicitly implemented and part of the programme code. No intelligence implemented, interaction is based on stimulus-response basis. Responsibility for	> Page 28 The HLGE solicits our partaking in the practical operationalization of the assessment list on four particular use cases of AI, selected based on the input from the 52 AI HLEG experts and the members of the European AI Alliance: (1) Healthcare Diagnose and Treatment, (2) Autonomous Driving/Moving, (3) Insurance Premiums and (4) Profiling and law enforcement. (1) Healthcare Diagnose and Treatment Bioethics requires special vigilance and attention to the announcement of the "bad news". If the diagnosis and prognosis have been established by an RN, care must be taken to ensure that doctors pay more attention to the way information is presented, based on the psychology of the patient and the humanity of the caregiver. (2) Autonomous Driving/moving Autonomous cars communicate with each other: car companies must ensure that shared data remains private and anonymous. In addition, transportation laws differ from member state to member state. It seems necessary to work on the harmonization of transport rules or at the very least on the road signing so that autonomous cars can operate optimally and without damage risks. (4) Profiling and Law Enforcement The burden of proof is a part of the rule of law in the EU. AI based law decisions must not invert the burden of proof principle and place it on the user, should the AI-based law decision be contested.	FEEDBACK DRAFT AI ETHICAL GUIDELINE As a preamble, we would like to acknowledge the quality of the draft produced by the HLEG on AI and to reckon that we are off to a very good start. This draft seems to be a powerful basis to which we, Professor Nathalie Nevejans, Robotics and Artificial Intelligence Law and Ethics Expert, Lecturer in Law, University of Artois (France), Member of the CNRS Ethics Committee, Expert to the European Parliament and Laetitia Pouliquen, Director, NBIC Ethics, are honored to contribute. First, we are truly grateful that our recommendation on Ethics to the European Commission for more transparency on the impact assessment progress of the Machine Directive 2006/42/CE was taken into account. This impact assessment leading to possible changes in regulations on the critical issue of machines and algorithms liability is now possible via the Machinery Directive revision feedback until next February (https://ec.europa.eu/info/law/better-regulation/initiatives/ares-2018-6426989_en). Second, the make-up of the HLGE of 52 experts remains unbalanced: with an overwhelming number of industry and federations stakeholders, we stress out the rare number or absence of philosophers, ethicists, religious leaders, anthropologists, consumer organizations and health experts. An enhanced AI HLEG would better guarantee the necessary respect for human rights based on a deep human-machine understanding. Please refer to our previous post with our Ethical recommendations on AI published in December 2018. Third, it is noteworthy that the Oviedo Convention was not adopted by all members of the Council of Europe. Moreover, even among the signatories, several nations and states took a very long time before ratifying the Convention. Consequently, the question we ask is “Will these AI ethical guidelines turn into soft law to ensure the cultural shift needed for AI developers to take onboard these ethical constraints?” Fourth, another thought-provoking question remains: why should we ask for a willing base adoption when human rights are de facto overarching rights? Should we sign on rights that are already defined in the EU Charter of Fundamental Rights and the European Convention on Human Rights, be it for AI or anything else? However, to ponder this remark, the technical variations of the EU ethical principles are novel and need to be designed. Fifth, ‘Trustworthy AI’ could also be branded as ‘Ethical Inside’ (like the famous and efficient ‘Intel Inside’) with an ISO accreditation.

Algorithmic scoring and notation impact very large areas that are not by definition limited: they are bound to have a "snowball" effect. Credit scoring impacts access to housing, to employment, to proper schooling etc.
Second, as already mentioned, by excluding oneself from the AI-based algorithms, he/she shall not benefit from what he or she claims, subject to the evaluation of the algorithm (eg: a bank credit).
Please add facial recognition for commercial use as strictly forbidden for “ethical integrity”.

In 5.5 Potential longer-term concerns :
> Page 12 ADD Human-Machine responsibility
We note that the Joint Research Center report (https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/artificial-intelligence-european-perspective) - Artificial Intelligence: A European Perspective JRC rightly recalls the principles and EU values impacted by AI : autonomy, identity of individuals, dignity and right to privacy and personal data protection.
The JRC paragraph on dignity draws our attention to the possible erosion of human rights:
“Individuals’ rights and responsibilities could start eroding as a result of the increasing interaction of humans and machines (EDPS, 2018). At the moment, smart devices have no moral responsibility and that is why it could be potentially harmful to let them manage human beings (EGE, 2018). However, the European Parliament called for the EC to consider a specific legal status for robots (EP, 2017), which is still a controversial proposal when considering, for instance, that at the present time accountability is ultimately related to human responsibility (EECS, 2016).”
First, it is inconceivable that individuals’ rights and responsibilities should erode as a result of the increasing interaction between humans and machines.
Second, the fact that, “at the moment, accountability is ultimately related to human responsibility” (EESC) is a good thing. However, we reiterate that the creation of a specific legal status for robots would be the wrong response the liability problem, as expressed in our Open Letter to the European Commission on AI and Robotics (http://www.robotics-openletter.eu/). Signed by 285 EU experts in AI, ethics and law, the signatories hereby affirm that the creation of a Legal Status of an “electronic person” for “autonomous”, “unpredictable” and “self-learning” robots is inappropriate from a technical, ethical and legal perspective.
Humans must always be responsible for their algorithms and for any damages caused. In fact, The European Group of Ethics of Science and Technology denies any moral standing to AI systems or robots in its report from March 2018 Artificial Intelligence, Robotics and ‘Autonomous’ Systems (http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf). EGE shares its moral reflections against the principle of autonomy. AIs are algorithms and robots are machines: “Human beings ought to be able to determine which values are served by technology, what is morally relevant and which final goals and conceptions of the behaviour lies with the developer. (2) Machine can learn and adapt but works on implemented/ given domain model; responsibility has to be with the developer since basic assumptions are hard coded. (3) Machine correlates internal domain model with sensory perception & information. Behaviour is data driven with regard to a mission. Ethical behaviour can be modelled according to decision logic with a utility function. (4) Machine operates on a world model as perceived by sensors. Some degree of self-awareness could be created for stability and resilience; might be extended to act based on a deontic ethical model. (5) Machine operates on a world model and has to understand rules & conventions in a given world fragment. Capability of full moral judgement requires higher order reasoning; however, second order or modal logics are undecidable. Thus, some form of legal framework and international conventions seem necessary and desirable. Systems that operate at level 4 can be said to have “Operational autonomy”. I.e., given a (set of) goals, the system can set its actions or plans."
With regard to Footnote 24, the definition, and especially the legal consequences for autonomy, lack a nuance and finesse, and are very obscure concerning points 3 to 5. It is not possible to approach as succinctly the questions of civil liability on such subtle points of distinction. And above all, these words suggest that the HLEG on AI believes that in this case, it is someone other than the developer who should be responsible, that is to say the machine itself. It would therefore be necessary to delete the legal references or modify the text.

Here is our proposal for deletion in Footnote 24:
"24. AI systems often operate with some degree of autonomy, typically classified into 5 levels ; as autonomy increases, the determination of the responsible person may be more difficult: (1) Domain model is implicitly implemented and part of the programme code. No intelligence implemented, interaction is based on stimulus-response basis. (2) Machine can learn and adapt but works on implemented/ given domain model; (3) Machine correlates internal domain model with sensory perception & information. Behaviour is data driven with regard to a mission. Ethical behaviour can be modelled according to decision logic with a utility function. (4) Machine operates on a world model as perceived by sensors. Some degree of self-awareness could be created for stability and resilience; might be extended to act based on a deontic ethical model. (5) Machine operates on a world model and has to understand rules & conventions in a given world fragment. Capability of full moral judgement requires higher order reasoning, however, second order or modal logics are undecidable. At Levels 4, but especially 5, a legal framework and international conventions seem necessary and desirable. Systems that operate at level 4 can be said to have “Operational autonomy”. I.e., given a (set of) goals, the system can set its actions or plans."

In 1. Requirements of Trustworthy AI 8. Robustness :
> Page 17

good are worthy to be pursued. This cannot be left to machines, no matter how powerful they are. […] Moral responsibility, in whatever sense, cannot be allocated or shifted to 'autonomous' technology."
Similarly, UNESCO's World Commission on the Ethics of Scientific Knowledge and Technology (COMEST) confirms their ban on a legal status for AIs and robots in its report of COMEST on robotics ethics calls the possible creation of a legal status for robots as: "highly counterintuitive to call them 'persons' as long as they do not possess some additional qualities typically associated with human persons, such as freedom of will, intentionality, self-consciousness, moral agency or a sense of personal identity" (Article 201, https://unesdoc.unesco.org/ark:/48223/pf0000253952)

> Page 12 ADD - Holistic view over AI
AI and Robotics are artefacts that could impact our humanity. Nevertheless, the line between human and machine must be unequivocally affirmed. Therefore, AI needs to be viewed holistically due to NBIC technology convergence: Nano, Bio, Information and Cognitive technologies will all use AI algorithms and interact with people, either externally or internally. Boundaries between restorative and augmentative health technologies Eg of Neurosciences
The European Commission should set the line between restorative and augmentative technologies and decide whether augmenting human beings using AI is acceptable or not, with regards to EU values. We believe that promoting an augmented humankind would increase unfairness and inequality and lead to a loss in individual s' rights in a democratic EU society.
As an example, let us pick the interaction between AI and neuroscience. Swiss researchers Marcello Ienca and Roberto Andorno give a relevant example (Ienca M. and Andorno R. Towards new human rights in the age of neuroscience and neurotechnology. Life Sciences, Society and Policy. 2017. Vol.13, n°1:5). Their research led them to believe that human rights in the age of neuroscience and neurotechnology are subject to four major threats:
• Right to cognitive freedom as the right to alter one's mental state by technical means and the right to refuse to do so. It is in fact the right not to be pressured to reveal data.
• Right to mental privacy as the right to prevent illegitimate access to our brain information - This is in fact the question of neuromarketing.
• Right to mental integrity as the right of individuals to protect their mental dimension from any potential danger, for example from hacking by a neural device (hacking of a neuro-device)
• Right to psychological continuity as the right to preserve one's personal identity and consistency of the individual behavior against unacceptable changes, even if the changes introduced are not per se dangerous.
Their research illustrates how, as AI is now used in all neurological technology, its impact has the potential to challenge who we are as humans. AI and robotics need to be considered holistically and not just as one element of a more global use of NBIC technologies.

In matter of reliability and resilience to attack, we should mention the obligatory measure of a "kill-switch" button to AI automated robots.

In 1. Requirements of Trustworthy AI 9. Safety :
> Page 18
"Moreover, formal mechanisms are needed to measure and guide the adaptability of AI" .
What does this exactly mean?

In 2. 1 Technical methods "Ethics & Rule of law by design (X-by-design)" :
> Page 19
"This also entails a responsibility for companies to identify from the very beginning the ethical impact that an AI system can have, and the ethical and legal rules that the system should comply with".
We perfectly agree. Nevertheless, it should be noted that it is very difficult to identify all the ethical impacts "from the very beginning". If all ethical impacts cannot be identified immediately (due to the novelty of the product, for example), they must be able to be identified later (after an experience of the product on the market and/or its varieties of use, for example).
In addition, this question raises the problem of whether only companies should be concerned. We believe that both States and the Public authorities should equally be concerned and responsible in the matter and should also be encouraged or obliged to endorse them.

"Explanation (XAI research)" :
"[…] In addition, sometimes small changes in some values of the data might result in dramatic changes in the interpretation, leading the system to confuse a school bus with an ostrich for example. This specific issue might be used to deceive the system".
It seems very relevant to give an example. There should be more in the guidelines.

						Is allowing the augmentation of humankind with AI or else, acceptable according to the EU values? We still need to address these philosophical and anthropological questions: what are the boundaries between human intelligence and artificial intelligence? Between humans and physical machines? Between natural life and Artificial life? The more we know about AI, the more it calls for a profound reflection on the boundaries between human intelligence and artificial intelligence. Determining what defines us as human beings will avoid the blurring between natural life and artificial life. Without this reflection, the questions of EU values and human rights would be irrelevant. The lines between restorative care and augmentation of humans need to be set. The European Commission should decide whether augmenting humankind using NBIC technology is acceptable to EU values. Investing significant EU funds for human restorative care technologies is desirable. However, we believe that augmenting humankind would result in unfairness and inequality. Individual rights would be more difficult to guarantee, even in a democratic society.		
Jacob	Turner	Author of Robot Rules: Regulating Artificial Intelligence	I agree with the focus in the Guidelines on trustworthiness, split into an ethics and a technical feasibility component. As I argue in Robot Rules: Regulating Artificial Intelligence (Palgrave Macmillan, 2018) (hereafter "Robot Rules"), setting ethical regulations for AI has two components: the political and sociological challenge of determining which ethical principles to use, and then the technical challenge of implementing these (see Chapter 8).However, the Commission may perhaps have missed some of the other novel problems which AI causes (and which are related to the overall question of trustworthiness). As I argue in Chapter 1 of Robot Rules, AI gives rise to three issues: 1. Responsibility – who is liable if AI causes harm, and who is the owner/ beneficiary if AI creates something of value?2. Rights – are there grounds for giving corporate personality to AI?3. Ethics – how should AI make choices, and are there any decisions it should not take?As to Responsibility (Issue 1), where AI systems make choices, there is no legal framework for determining who or what should be held responsible. It could be the programmer, owner, operator, a combination of the above, or perhaps none.Two features of AI make it difficult to hold the original programmer always responsible. First, AI is becoming more independent; some AI systems are now able to develop new AI. Secondly, the barriers between designers and users are being broken down as AI becomes more user-friendly.	Respecting fundamental rights is an appropriate aim.However, the first two suggested principles of "Beneficence" and of "Non maleficence" are so banal and vague as to be meaningless in practice. Their presence in the document detracts from its otherwise generally sensible approach. No one considers that Google's motto of "Don't be evil" constrains the company's actions (consider Project Maven – assisting DARPA and Project Dragonfly – assisting the Chinese Government; executives clearly considered both to be consistent with the motto). No party – whether it is a corporation or a national government – would ever consider itself to have breached these principles. I therefore suggest that the Commission removes them from the next iteration of the document, or otherwise risk any contribution being dismissed by commentators as merely empty words and rhetoric. The right to autonomy is described as a right to an "opt out". However, with regards to the "Right to Object to Automated Processing" in GDPR Article 22, the Article 29 Working Party has declared that this amounts to an outright ban on decisions made solely by automated processing. This seems to go beyond the High-Level Expert Group's proposals. The prospect of always having a "human in the loop" is, in my view, highly problematic. As I explain in Chapter 8 of Robot Rules, this would resemble the 19th Century "Red Flag Laws" in the UK, where a person was required to walk in front of every	The list of ten requirements in the draft Guidelines is appropriate. As to transparency, the Commission should be very careful to ensure that requirements do not go beyond what is technically feasible, and also take into account that much human decision-making is not truly transparent. Rather than focussing on transparency for its own sake, it would be preferable to seek transparency focussed on correcting errors. Articles 13 and 14 of the GDPR will need to be interpreted sensitively in this regard or otherwise risk a trade-off between the effectiveness of certain systems which may do great good, with the ability to explain how they have made their decisions (see e.g. Edwards and Veale, "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For", 16 Duke Law & Technology Review 18 2017). The High-Level Expert Group could play a valuable role in setting guidance on these provisions at an early stage, or otherwise risk that they are given a very expansive interpretation by the CJEU as and when it is called upon to consider them. See Chapter 8 of Robot Rules for further discussion.A further mechanism to realising trustworthy AI might be to require that AI engineers/ data scientists become professionalised and regulated, in the same way as doctors, lawyers, architects and other professions. The EU could set overall standards and become a world leader in this regard. See Chapter 7 of Robot Rules for	I agree with the content of this section. Ideally it should be built into a full protocol which can then be implemented by businesses, governments and other deployers of AI systems. This is exactly the type of granular guidance that is needed (rather than nebulous principles such as "Beneficence").Beyond the assessment of trustworthy AI, the next step is to enforce this. The Commission should consider using its impressive array of legislative tools (regulations, directives, guidance etc), as well as working with member state governments, to put in place structures which encourage and even compel compliance with the relevant principles. As Paul Nemitz convincingly argues, AI impact assessments could present a valuable tool in this regard – drawing on the type of exercise which is already familiar both in terms of other computer technology and indeed wider domains (P. Nemitz,Constitutional democracy and technology in the age of artificial intelligence, Philosophical Transactions Of The Royal Society A: Mathematical, Physical And Engineering Sciences, 15 October 2018, https://doi.org/10.1098/rsta.2018.0089).. The AI Now Institute has prepared a draft impact assessment (https://ainowinstitute.org/aiareport2018.pdf), and IBM have also produced a useful suite of programs and guidance in this regard (https://www.research.ibm.com/artificial-	Paul Nemitz is correct to call for "a new culture of incorporating the principles of democracy, rule of law and human rights by design in AI" (Nemitz, Op Cit).Before writing rules, the Commission should give significant consideration to the prior stage, of designing the institution(s) capable of writing those rules and then enforcing them. In short, it is important to ensure democratic legitimacy and public understanding for such a body, as a preliminary to the development of any rules. Tools for increasing public engagement will vary from country to country but useful methods might include wide-ranging public consultations undertaken by national and regional governments (as opposed to a centralised one such as the present stakeholders' consultation, which necessarily reaches fewer parties). Chapters 6, 7 and 8 of Robot Rules provide a roadmap of the steps that the EU (as well as national governments) can take in order to build institutions that can govern AI effectively. In order for the Commission to achieve its laudable aim of creating trustworthy AI, it needs to go much further in an effort to reach constituents. Large companies are already well-represented on the High-Level Expert Group, but small to medium sized enterprises, academia, and wider civil society much less so. National and regional governments should be mobilised in this effort. The danger is that without doing so, any principles may either be ignored or they will

friendly. In addition to harm, most legal systems at present do not tell us who is responsible if AI creates something beneficial, which might be covered by intellectual property protections had it been done by a human. For instance, AI art is already becoming very desirable. Current Intellectual Property laws (in the EU and elsewhere) make insufficient provision for the protection of AI creativity. One solution to the question of responsibility is to give AI its own legal personality (Issue 2). This would involve giving an AI system the ability to hold property, to sue and be sued. Indeed, the European Parliament proposed legal personality for AI in its resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)). Many "AI Experts" rejected this idea out of hand, but the vast majority of the arguments that they made against AI legal personality could equally be made against any legal persons, including companies. Artificial legal persons are not new; we have had limited liability companies with their own legal personality for thousands of years, and they play a valuable economic role.It will only take one country to start a domino effect for AI personality. If a smaller jurisdiction with the ability to move quickly – perhaps Singapore, Malta or the British Virgin Islands – adopts AI legal personality, then others are likely to follow so as not to lose out on any competitive advantage.EU countries are already required to recognise other (profit-making) corporate persons registered in other member states (see Art 54 TFEU). As such, the Commission should give serious consideration to the question of whether legal personality for AI might be desirable, and if so, how this would be structured. If it does not, the risk is that a country inside or even outside the EU will move first on this, and the EU will become a rule-taker rather than a rule-maker in this regard.The question of legal personality for AI is explored in detail in Chapter 5 of Robot Rules (Chapter 4 addresses rights for AI from a moral perspective)As to ethics (Issue 3), there are two main issues: (i) how should AI make decisions? and (ii) are there any decisions AI should not take?The current draft guidelines engage somewhat with both of these issues, but splitting the two would assist in clarifying the analysis. In my view there are not necessarily any correct answers to either question (i.e. I take a positivist approach). Rather (as expanded upon below in my general comments), the initial focus should be on designing institutions which are capable of consulting the relevant populations to find the answers.

automobile waving a red flag, so as to warn pedestrians and other road users. It may have been effective in doing so, but it neutralised the effectiveness of automobiles in driving any faster than walking pace. This is thought by some economic historians to have had a highly-deleterious effect on the UK's nascent car industry (and handed the advantage to the US). The EU risks doing the same if it insists on inserting a human in the loop. Having a "human on the loop" is a much better approach in terms of balancing accountability and autonomy with the vast benefits that AI can bring. Such arguments are expanded on in Chapter 8 of Robot Rules.As to autonomous weapons, I am pleased to see that the High-Level Expert Group has taken a balanced approach in recognising to their potential usefulness in saving civilian lives. As I have argued elsewhere (see https://blogs.spectator.co.uk/2017/08/we-should-regulate-not-ban-killer-robots/) an outright ban on autonomous weapons systems is likely to be both ineffective and counter-productive. What is needed is better regulation of such systems, so that as and when they are used, they are able to fulfil the laws of warfare (and indeed to exceed the abilities of humans to do so).

further description of why this would be helpful and how it could be achieved. In addition, and as proposed by the EU Parliament in February 2017, it would also be useful for individual users of AI (who are not specialist) to undergo some form of training (similar perhaps to the acquisition of a drivers' license). See again Chapter 7 of Robot Rules.

intelligence/trusted-ai/). Fostering a culture of compliance is important, but without any coercive measures to support this (or at least some parts thereof) then the ethical guidance promulgated may end up being ignored. The EU is uniquely well-placed to be a world leader in the creation of binding legislation in due course; it should seize this opportunity before others do.

be seen as an elitist project and therefore rejected. We can see from the UK's recent experience with the EU that a broadly very positive set of rules and institutions (such as those of the EU) may be spurned by a population if it does not feel sufficiently involved in the law-making process. It is imperative that we avoid this outcome for the EU's AI ethical principles.

Sergio	Sestili	myself (https://www.linkedin.com/in/sergio-sestili-78a6a714/)	I totally agree with the principles and values expressed as foundation for AI Systems development. I'd like to add another aspect, in my opinion, as important as the ethical purposes: the social aspect of introducing more and more AI systems in the human communities. The more we go forward in the future, the more the AI systems will be increasingly complex and able to perform more difficult tasks with autonomy, precision and reliability even greater than an human operator. Several statistics shows that in the next 5-10-15 years, AI systems will be able to make a significative percentage of jobs that now are performed by humans, to be optimistic 40%, and even more. "To be able to" does not mean that humans will automatically loose their jobs because of AI, but I believe that it should be imperative to build a legal framework of lows and rules to manage this scenario, in order to protect the human-side of the human-centric-AI. I've always been enthusiastic for science and technology, and for AI in particular. And I think that the raise of AI it is simply an inevitable process. Having no specific rules may inevitably lead to fall in a far-west-like scenario, where, if driven only by increasing profits, entrepreneurs would prefer to fire humans and hire AI systems. Because we all agree the main principles that AI must be human-centric, ethic, "do no harm" humans, I see that we should intervene as soon as possible to guide AI usage in a path inside those principles. There will be a certain dose of inevitability in having AI that will substitute humans, perhaps this process will accelerate as we go deeper into the future. But even an inevitable process must be guided as much as we can.	Our civilization is based on having conquered rights and duties, guaranteed by laws. It may be a good idea for me to use the same approach with AI systems, and put in place a legal framework, a set of specific rules to guide this process. And by consequence AI Systems must be built to respect all the those rules. In addition, where an AI system si going to substitute humans, those humans should access to a reconvert path that let them not to lose economic income for continuing their life path. This will contribute to enforce trustworthy in AI. I agree that a fundamental aspect in actual and future AI systems is Data Governance. Maintaining the expected level of trustworthy means in terms of Data to have a strong attention to the data injected in AI systems, not only during initial training but during all their lifecycle. In order to mitigate bias and avoid intentional bad data injection, it is necessary to have some sort of certification path for data used to train and re-train AI systems. The long term goal should be to request that for using AI systems in particular domains, they should acquire the above certifications. An an example let's consider all the certification steps needed for an aerospace startup company to certify his own rocket to be able to fly astronauts for NASA. There must be a dedicate AI certification authority devoted to assess and give the AI-Certifications. Another relevant aspect should be related to AI systems that work in an human environment. I would consider the hypothesis to impose them to obey at minimum the same laws that every human must respect. Hence, AI systems must be build having in mind that they must obey humans laws (do not cross a street if light is red). I believe that AI and humans should have same rights and duties, not to having unfair competition, especially in a future where Ai systems will be more and more advanced. It means that, for example, if an human driven taxi can not work more than 8 hours continuously, the same should be for an AI-taxi. This much more in the beginning of this age. In a more distant future where all taxi will be guided by an AI System, this constraint could be removed, producing non unfair competition with humans "colleagues" but offering a better service for the community. And in a medium/distant future where AI system could substitute great part of humans, humans should be given for free enough finance to live, even without working. Experiments of this Universal Income Concept seems to be encouraging, and this may be a long-term-goal to analyze and hopefully reach, freeing humans from jobs that can be dona by AI and letting them to express their deep potentials. I see this as a higher civilization level, the greatest achievement we could aspire to, together with all the fundamental principles that are the foundation of our community.	I thing that putting in place a legal framework for managing AI is a fundamental action to be taken. In addition should be defined a set of tools to be used to evaluate AI systems too. I'd think about the possibility to to evaluate AI Systems by using special AI Systems. Within open source world everyone, even criminals, can implement AI systems. We must put a borderline. Only certified AI systems must be allowed to operate in critical environments, infrastructures, contexts.	I find extremely positive and fundamental to have this document as a foundation for EU approach to manage this exponential technology. I agree in having a global approach because physical borders are of no meaning in our tech times. Most of AI systems we are currently using comes from outside Europe. And I like the idea of an Europe that can define rules that may become a successful case study for the global community. World is becoming smaller and smaller and one of the powers of AI should be to positively contribute to that, for example to bringing people together helping to overcame the languages barriers. But this "singularity" we are living in, must be managed to achieve our goals according our main principles, for common good.

Martin	Haimerl	Hochschule Furtwangen University			The principles of “Do Good” and “Do not harm” cannot be guaranteed in a strict way, in general. For example, a medical device can be very beneficial for a lot of patients. But if something goes wrong, it can also be detrimental (in the worst case, lethal) in other cases. Thus, it would make sense to sum up with a requirement regarding a proper relationship between risks (“do not harm”) and benefits (“do good”), i.e. the requirement for assessing this relationship during the development process (and keeping it updated during the entire product life cycle). See also my comment to chapter II regarding a requirement to perform a risk-benefit analysis.	Ad 3: Design for all A design-for-all approach does not work in all situations. E.g. for medical devices, it is crucial (and a requirement) that you define a user population, which is allowed to use/apply the product. This a major requirement w.r.t. safety. A medical device can only be used by users who are sufficiently skilled w.r.t. the usage of the product (e.g. experienced surgeons or other doctors). Using the products without sufficient background knowledge is dangerous. This also applies to apps which use AI e.g. to suggest specific treatments/drugs or provide other recommendations for improving the health status. Ad 5: Non-discrimination For some disciplines, an AI based product has to prefer certain groups. For example, some medical products work better for female than for male patients, or vice versa. A straight non-discrimination is not valid in such cases. This discrimination does not appear unintentionally, but is at the core of the particular application. A strategy to define non-discrimination can be very challenging in these situations. Ad 9: Safety In the Executive Summary (p. i), you emphasize that risks should be balanced / outweighed by benefits. (“Given that, on the whole, AI’s benefits outweigh its risks, we must ensure to follow the road that maximises the benefits of AI while minimising its risks.”) However, I did not find a realisation task which addresses this major requirement. Shouldn’t some kind of risk-benefit analysis/report be included? For medical devices, such an analysis is required. This could also help to enforce that AI developers analyse / explicitly address the relationship between risks on the one side and benefits on the other side. From my perspective, section 9 in this guideline is very short in comparison to its importance. For medical devices, i.e. in the EU Medical Device Regulation, the risk-benefit analysis is a cornerstone on the way to release a product. Additional topic: I also miss a requirement for defining the basic goal and context which a system addresses. A system can only be analysed regarding risks and benefits, if the goal and context of its application are clear. The same applies to the validation of the system. AI is often not a stand-alone software tool, but it is integrated into an environment. It acts in this environment and interacts with it. Environment may refer to technical as well as social integration of the system. It is basically related to the application which it supports. If the application is not specified, the definition of a risk e.g. does not really make sense. Risks are usually related to harm and harm can usually only be addressed on an application level (and not on a level of a part/module of the system). In many cases, the defining system will not be the AI system, but the application behind, e.g. a medical device with a specific purpose. A requirement to specify this goal / context and dedicated applications which are intended to use the AI system would be crucial, from my point of view – at least for some domains / applications.		Basically, I clearly support the approach to set up guidelines / regulations to guide and support the development of AI systems. Currently, many areas of AI systems seem to work in an almost completely unregulated environment. However, I would suggest including more people who have direct experience with regulations in highly regulated domains like automotive, aerospace, or medical devices. At least for medical devices, it seems that not many of them have been included (though I did not analyse this in detail). Some parts of the documents are on a substantially different line compared with Medical Device Regulation, which is the main regulation in this sector within the EU. For example, this refers to the missing risk-benefit conclusion, dedicated user population, or application context in general for a product.

Anonymous	Anonymous	Anonymous		Regarding the potential long term concerns (subsection 5.5), if any should have even a minimal risk of happening then regulators must take them into consideration even if only at a basic level at this stage.			
Christian	Freksa	University of Bremen		3.4 Equality, …"Equality means equal treatment of all human beings, regardless of whether they are in a similar situation. Equality of human beings goes beyond non-discrimination, …"It is very important not to shorten "equality of treatment of human beings" to "equality of human beings". We must give equal treatment to all human beings despite the wonderful diversity, i.e. inequality, of human beings. Diversity is mentioned at various places in the document correctly as a feature of European culture. Thus, "Equality of human beings" => "Equal treatment of human beings".p.10, Principle of Justice"the positives and negatives resulting from AI should be evenly distributed"Misleading statement. Could be read as "the more good AI applications we have, the more evil applications we can permit".p.11, Critical concerns"A balance must thus be considered between what should and what can be done with AI,""balance" is misleading, as what should and what can be done are not on the same level and therefore cannot be balanced. Presumably, what should be done is a subset of what can be done.		A good start!The Executive Summary, par. 3 states:"Given hat, on the whole, AI's benefits outweigh its risks" This is the goal, not the given. Many would not agree with this (prejudiced) assumption."we must ensure to follow the road that maximises the benefits of AI while minimising its risks."This statement holds independently of the assumption in the beginning of the sentence.
Anonymous	Anonymous	Anonymous	Wenn die Europäische Kommission zu ihrem Entwurf ihrer Ethik-Leitlinien für Künstliche Intelligenz zur Stellungnahme aufruft, wie in ihrer Pressemitteilung vom 12.12.18 und erneut am 15.1.2019 geschehen, ist nicht erklärlich, weshalb die konkrete Konsultation sich auf den Standpunkt stellt, dass es sich bei den Richtlinien (!) nicht um Standpunkte der Europäischen Kommission, sondern eines Expertengremiums handele. Hier erwarte ich als EU-Bürgerin keine Flucht ins Unverbindliche, sondern Übernahme der Verantwortung für – notwendige – Richtlinien zum ethischen Umgang mit künstlicher Intelligenz. „The Guidelines are addressed to all relevant stakeholders developing, deploying or using AI, encompassing companies, organisations, researchers, public services, institutions, individuals or other entities. In the final version of these Guidelines, a mechanism will be put forward to allow stakeholders to voluntarily endorse them." „A mechanism will be put in place that enables all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis. This will be set out in the final version of the document." Welchen Sinn hat bei der Entwicklung ethischer Prinzipien die Freiwilligkeit deren Beachtung? In einer Realität, die durch ausgeprägte Asymmetrie der Einflussmöglichkeiten gekennzeichnet ist – Verbraucher auf der einen Seite mit praktisch kaum einer Wahl – auf der anderen Seite große Unternehmen, deren Algorithmen als Geschäftsgeheimnisse geschützt werden (vgl. die Rechtsprechung des BGH zur Schufa, die die Algorhythmen zu schützenswerten Geschäftsgeheimnissen erklärte) halte ich das Abstellen auf	Auf S. 7 heißt es: „Citizens should never be subject to systematic scoring by government." Dieser Aussage stimme ich in vollem Umfang zu, vor allem, wenn ich an die Entwicklungen in China denke, wo intensiv an einem Bürger-Scoring gearbeitet wird. Im gleichen Atemzug frage ich mich, weshalb diese Einschränkung nur für Regierungen gelten soll. Aus meiner Überzeugung darf es auch Unternehmen nicht erlaubt werden, den Bürger gläsern zu machen, in dem Scoring dazu führt, dass alle verfügbaren Daten – ohne dass der Einzelne die realistische Möglichkeit hätte, deren Richtigkeit zu prüfen – für Kreditprüfungszwecke, Entscheidungen über Arbeitsverhältnisse, Versicherungen u.ä. verwendet werden. In die gleiche Richtung geht die Argumentation auf S. 17, wenn es heißt: „Transparency is key to building and maintaining citizen's trust in the developers of AI systems and AI systems themselves. Both technological and business model transparency matter from an ethical standpoint. Technological transparency implies that AI systems be auditable,14 comprehensible and intelligible by human beings at varying levels of comprehension and expertise. Business model transparency means that human beings are knowingly informed of the intention of developers and technology implementers of AI systems." Auch hier reicht es nicht, auf die Prüffähigkeit von Algorhythmen abzustellen, wenn nicht gleichzeitig sichergestellt wird, dass es unabhängige Prüfinstanzen gibt, die Algorhythmen auch prüfen können (und es nicht wie im bereits zitierten BGH-Urteil unter Verweis auf Geschäftsgeheimnisse verwehrt wird). Hier wäre wenigstens eine Kontrolle eine Art „TÜV" von Nöten, um	Ins Leere werden auch Verhaltensregeln laufen, wie auf S. 22 „Codes of Conduct" formuliert. Als Verbraucherin habe ich keine Möglichkeit, zu kontrollieren, was z.B. bei den Regeln des autonomen Fahrens programmiert wurde, wenn die Entscheidung zwischen der Gefährdung von Insassen und von Passanten zu treffen ist.	Anmerkungen zum Verfahren Wenn die Europäische Kommission eine breite Partizipation ihrer Bürgerinnen und Bürger wünscht, hätte sie die gleichen Schritte einleiten können wie bei der Beteiligung der Bevölkerung an der Meinungsbildung zur Abschaffung der Sommerzeit. Zur Erinnerung: 5 Millionen Menschen haben daran teilgenommen. Wenn es der Europäischen Kommission nicht gelingt ein Papier, das mehr als 20 Seiten umfasst, aber ein so wichtiges Thema wie den ethischen Umgang mit künstlicher Intelligenz umfasst, nicht in den Amtssprachen der EU veröffentlicht, lässt dies nur den Schluss zu, dass eine gründliche Befassung der Bevölkerung nicht erwünscht ist. Dieser Eindruck drängt sich auch bei der Lektüre derjenigen auf, die in dem von der Kommission berufenen Gremium sitzen und den Entwurf der Richtlinien verfasst haben. Was haben IBM und Google als US-amerikanische Großkonzerne mit der ethischen Verantwortung innerhalb der europäischen Union zu tun? Warum fehlen in dem Gremium kritische Fachleute wie z.B. aus dem Chaos-Computer-Club? Ebenso wenig überzeugt der Zeitpunkt der Öffentlichkeitskampagne, die zunächst auf einen Monat begrenzt wurde und am 18.12.18, also kurz vor den Weihnachtsfeiertagen begonnen wurde und den Schluss nahe legt, dass man nur eine pro-forma Beteiligung macht. An diesem Eindruck ändert auch die minimale Verlängerung der Beteiligungsfrist nichts. Vielmehr erweckt der erzeugte Zeitdruck unangenehme Erinnerungen an das Verfahren im Trilog-Verfahren bei der Datenschutzgrundverordnung. Auch dort wurde ein künstlicher Zeitdruck aufgebaut,

			Freiwilligkeit für reine Alibitätigkeit. Ich halte es für höchst unwahrscheinlich, dass durch Einhalten der Richtlinien Verbraucher in die Lage versetzt würden, die Tragweite ihrer Handlung bei einer Einwilligung zu ermessen. Schließlich fehlt es den Verbrauchern auch an Möglichkeiten zur Kontrolle. Das entspricht – anders als auf S. 2 dargestellt – gerade keiner wirklichen effektiven Willensausübung.	ethische Prinzipien bei der Verwendung der Algorhythmen zu ermöglichen. Verbraucher könnten dann auch ihre Produkt/Dienstleistungsentscheidung an Hand des „TÜV-Siegels treffen. Ohne neutrale Kontrolle bleibt es bei der außerordentlich asymmetrischen Rolle des Verbrauchers /Bürgers einerseits und der Stelle, die hinsichtlich ihrer Algorhythmen keine Kontrolle befürchten muss.	der im Ergebnis dazu geführt hat, dass die DSGVO zahlreiche handwerkliche Mängel enthält und die DSGVO an der Marktasymmetrie, die im Internet herrscht, gerade in Bezug auf die Großen der Branche wie Google, Facebook, Amazon nicht wirklich etwas geändert hat. Heutzutage können Sie nicht einmal mehr ein Antivirusprogramm kaufen, ohne dass der Anbieter von Ihnen verlangt, dass Sie Ihre Daten zu Tracking-Zwecken zur Verfügung stellen. Als Verbraucher haben Sie keine wirkliche Wahl, weil es eben keine Antivirenprogramme gibt, die auf Tracking verzichten. Art 7 der DSGVO läuft insoweit völlig leer. Es heißt zwar in den Vorbemerkungen zum Ethik-Richtlinienentwurf als Ziel der Richtlinien: „Provide, in a clear and proactive manner, information to stakeholders (customers, employees, etc.) about the AI system's capabilities and limitations, allowing them to set realistic expectations" Wenn das tatsächlich das Ziel ist, bedarf es einer erheblich größeren Beteiligung der Bevölkerung, um für Informationen und Diskussionen zu sorgen.
Rieks	op den Akker	emeritus ass. professor of the University of Twente, the Netherlands	Indeed, AI is one of the most exciting and challenging products of the modern western intellect. The value of the Guidelines offered as well as the chances of a successful implementation and operationalization thereof strongly depend on our understanding of AI as an expression of the modern way people stand towards nature, live, society, themselves. And this is how I would like to understand the "human-centric approach to AI" that is taken. AI is not simply a bunch of techniques. AI is a way of thinking, a product of the specific way western scientific culture has developed. AI is the expression and the product of the scientific way we conceive the world and shape and make sense of our lives, how we conceive our work, how we organize society, our education, security, and health. The important question IMO is the following: What are the values and ideas implicitly held by this way of thinking? What does it mean to be autonomous beings? Trust and responsibility are core concerns related to AI. And this is so as a consequence of the functional stance AI takes. This functional stance takes behavior, work, communication, interaction as a depersonalised proces. The person and its values are out of sight. The replacement of man by robots and other autonomous systems assumes this proces of robotisation and formalisation of our work and activities. Introduction of technology may never take away the responsibility of people. If we take away responsibility from a person we take away a fundamental quality. When we decide to take the human out-of-the-loop we are responsible for doing that. Machines, however intelligent they may be conceived, cannot be held responsible for what they "do". AI based on machine learning assumes that nothing really new can happen, that data		

and patterns learned from data covers everything that will ever happen in future. This means that medical decision support systems may never be seen as dictating what a medical surgeon should do in a specific situation. Such a system must always be transparant and be able to explain its behavior. Intelligence in the sense of the capacity to decide what to do best in a concrete situation is something that we cannot implement in general rules or statistical tables learned from data.

"the Guidelines hence do not aim to provide yet another list of core values and principles for AI, but rather offer guidance on the concrete implementation and operationalisation thereof into AI systems." In my opinion it does not make sense to talk about trustfull AI systems abstract from the organisation, the work, in which the technology functions. If we look at concrete applications of AI we always see conflicting interests of parties involved. Users want to know all ins and outs of the AI they use, but companies are not always willing to explain the algorithms (a well knows problem in predictive policing). Information is often withhold by parties to protect personal privacy or for political, economical or security reasons (see for example the discussion we had in the Netherlands about the new law on information and security - "de sleepwet") .

Conventional wisdom assumes that the police are in control of their investigative tools. But with surveillance technologies, this is not always the case. Increasingly, police departments are consumers of surveillance technologies that are created, sold, and controlled by private companies. These surveillance technology companies exercise an undue influence over the police today in ways that aren’t widely acknowledged, but that have enormous consequences for civil liberties and police oversight.
(Elizabeth Joh in: ``The Undue Influence of Surveillance technology Companies on Policing’’ (2017))

"it should be noted that no legal vacuum currently exists, as Europe already has regulation in place that applies to AI."
This ignores the accountability gap in the context of autonomous weapons (aka Killer Robots).
Solutions sought by attributing legal status to robots will not work.
``From an ethical and legal perspective, creating a legal personality for a robot is inappropriate whatever the legal status model.'' (see the OPEN LETTER TO THE EUROPEAN COMMISSION
ARTIFICIAL INTELLIGENCE AND ROBOTICS).

Anonymous	Anonymous	Anonymous	Wenn die Europäische Kommission zu ihrem Entwurf ihrer Ethik-Leitlinien für Künstliche Intelligenz zur Stellungnahme aufruft, wie in ihrer Pressemitteilung vom 12.12.18 und erneut am 15.1.2019 geschehen, ist nicht erklärlich, weshalb die konkrete Konsultation sich auf den Standpunkt stellt, dass es sich bei den Richtlinien (!) nicht um Standpunkte der Europäischen Kommission, sondern eines Expertengremiums handele. Hier erwarte ich als EU-Bürgerin keine Flucht ins Unverbindliche, sondern Übernahme der Verantwortung für – notwendige – Richtlinien zum ethischen Umgang mit künstlicher Intelligenz.„The Guidelines are addressed to all relevant stakeholders developing, deploying or using AI, encompassing companies, organisations, researchers, public services, institutions, individuals or other entities. In the final version of these Guidelines, a mechanism will be put forward to allow stakeholders to voluntarily endorse them.“ „A mechanism will be put in place that enables all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis. This will be set out in the final version of the document.“ Welchen Sinn hat bei der Entwicklung ethischer Prinzipien die Freiwilligkeit deren Beachtung? In einer Realität, die durch ausgeprägte Asymmetrie der Einflussmöglichkeiten gekennzeichnet ist – Verbraucher auf der einen Seite mit praktisch kaum einer Wahl – auf der anderen Seite große Unternehmen, deren Algorithmen als Geschäftsgeheimnisse geschützt werden (vgl. die Rechtsprechung des BGH zur Schufa, die die Algorhythmen zu schützenswerten Geschäftsgeheimnissen erklärte) halte ich das Abstellen auf Freiwilligkeit für reine Alibitätigkeit. Ich halte es für höchst unwahrscheinlich, dass durch Einhalten der Richtlinien Verbraucher in die Lage versetzt würden, die Tragweite ihrer Handlung bei einer Einwilligung zu ermessen. Schließlich fehlt es den Verbrauchern auch an Möglichkeiten zur Kontrolle. Das entspricht – anders als auf S. 2 dargestellt – gerade keiner wirklichen effektiven Willensausübung.	Auf S. 7 heißt es: „Citizens should never be subject to systematic scoring by government.“ Dieser Aussage stimme ich in vollem Umfang zu, vor allem, wenn ich an die Entwicklungen in China denke, wo intensiv an einem Bürger-Scoring gearbeitet wird. Im gleichen Atemzug frage ich mich, weshalb diese Einschränkung nur für Regierungen gelten soll. Aus meiner Überzeugung darf es auch Unternehmen nicht erlaubt werden, den Bürger gläsern zu machen, in dem Scoring dazu führt, dass alle verfügbaren Daten – ohne dass der Einzelne die realistische Möglichkeit hätte, deren Richtigkeit zu prüfen – für Kreditprüfungszwecke, Entscheidungen über Arbeitsverhältnisse, Versicherungen u.ä. verwendet werden.In die gleiche Richtung geht die Argumentation auf S. 17, wenn es heißt:„Transparency is key to building and maintaining citizen’s trust in the developers of AI systems and AI systems themselves. Both technological and business model transparency matter from an ethical standpoint. Technological transparency implies that AI systems be auditable,14 comprehensible and intelligible by human beings at varying levels of comprehension and expertise. Business model transparency means that human beings are knowingly informed of the intention of developers and technology implementers of AI systems.“Auch hier reicht es nicht, auf die Prüffähigkeit von Algorhythmen abzustellen, wenn nicht gleichzeitig sichergestellt wird, dass es unabhängige Prüfinstanzen gibt, die Algorhythmen auch prüfen können (und es nicht wie im bereits zitierten BGH-Urteil unter Verweis auf Geschäftsgeheimnisse verwehrt wird). Hier wäre wenigstens eine Kontrolle eine Art „TÜV“ von Nöten, um ethische Prinzipien bei der Verwendung der Algorhythmen zu ermöglichen. Verbraucher könnten dann auch ihre Produkt/Dienstleistungsentscheidung an Hand des „TÜV-Siegels treffen. Ohne neutrale Kontrolle bleibt es bei der außerordentlich asymmetrischen Rolle des Verbrauchers /Bürgers einerseits und der Stelle, die hinsichtlich ihrer Algorhythmen keine Kontrolle befürchten muss.	Ins Leere werden auch Verhaltensregeln laufen, wie sie auf S. 22 die „Codes of Conduct“ formulieren. Als Verbraucherin habe ich keine Möglichkeit, zu kontrollieren, was z.B. bei den Regeln des autonomen Fahrens programmiert wurde, wenn die Entscheidung zwischen der Gefährdung von Insassen und von Passanten zu treffen ist.		Die Europäische Kommission bietet Gelegenheit zur Stellungnahme. Wenn sie eine breite Partizipation ihrer Bürgerinnen und Bürger wünschte, hätte sie allerdings die gleichen Schritte einleiten müssen wie bei der Beteiligung der Bevölkerung an der Meinungsbildung zur Abschaffung der Sommerzeit. Zur Erinnerung: 5 Millionen Menschen haben daran teilgenommen. Wenn es der Europäischen Kommission nicht gelingt ein Papier, das mehr als 20 Seiten umfasst, aber ein so wichtiges Thema wie den ethischen Umgang mit künstlicher Intelligenz umfasst, nicht in den Amtssprachen der EU veröffentlicht, sondern nur in Englisch, lässt dies nur den Schluss zu, dass eine gründliche Befassung der Bevölkerung nicht erwünscht ist. Dieser Eindruck drängt sich auch bei der Lektüre derjenigen auf, die in dem von der Kommission berufenen Gremium sitzen und den Entwurf der Richtlinien verfasst haben. Was haben IBM und Google als US-amerikanische Großkonzerne mit der ethischen Verantwortung innerhalb der europäischen Union zu tun? Warum fehlen in dem Gremium kritische Fachleute wie z.B. aus dem Chaos-Computer-Club?Ebenso wenig überzeugt der Zeitpunkt der Öffentlichkeitskampagne, die zunächst auf einen Monat begrenzt wurde und am 18.12.18, also kurz vor den Weihnachtsfeiertagen begonnen wurde und den Schluss nahe legt, dass man nur eine pro-forma Beteiligung macht. An diesem Eindruck ändert auch die minimale Verlängerung der Beteiligungsfrist nichts. Vielmehr erweckt der erzeugte Zeitdruck unangenehme Erinnerungen an das Verfahren im Trilog-Verfahren bei der Datenschutzgrundverordnung. Auch dort wurde ein künstlicher Zeitdruck aufgebaut, der im Ergebnis dazu geführt hat, dass die DSGVO zahlreiche handwerkliche Mängel enthält und die DSGVO an der Marktasymmetrie, die im Internet herrscht, gerade in Bezug auf die Großen der Branche wie Google, Facebook, Amazon nicht wirklich etwas geändert hat. Heutzutage können Sie nicht einmal mehr ein Antivirusprogramm kaufen, ohne dass der Anbieter von Ihnen verlangt, dass Sie Ihre Daten zu Tracking-Zwecken zur Verfügung stellen. Als Verbraucher haben Sie keine wirkliche Wahl, weil es eben keine Antivirenprogramme gibt, die auf Tracking verzichten. Art 7 der DSGVO läuft insoweit völlig leer.Es heißt zwar in den Vorbemerkungen zum Ethik-Richtlinienentwurf als Ziel der Richtlinien: „Provide, in a clear and proactive manner, information to stakeholders (customers, employees, etc.) about the AI system’s capabilities and limitations, allowing them to set realistic expectations“Wenn das tatsächlich das angestrebte Ziel ist, bedarf es einer erheblich größeren Beteiligung der Bevölkerung, um für Informationen und Diskussionen zu sorgen.
Anonymous	Anonymous	Anonymous	Congratulation to the Commission's vision of setting up a system of trustworthy AI, made in Europe. Trustworthiness is the core problem of AI, citizens agree, having experienced misuse of private datas and being aware of AI made in China, the model of dictatorship by AI.	The detailed chapters (B I und II) listing all relevant varieties of rights, principles and values trustworthy AI should respect take the bigger part of the document, proving that the High-Level Expert Group knows why citizens don't trust AI. Convincing: Some negative aspects of AI are mentioned - AI which by its very nature is a continuous temptation for misuse.	See the above comment on Chapters I and II	The - dangerous - failure of the draft is the proposed way to implement these prophetic views of trustworthy AI into reality - it leaves this to "the process", a "circular model", aligned with "the spirit" of trustworthy AI. That means the implementation is left to persons, institutions and even machines, guiding the upspeeding development of AI, without public control!	The dangerous failure of the draft seems to be a consequence of composing the High-Level Expert Group: No representatives of European Churches, of European Trade Unions, of European Lawyer's Associations, of groups working on ethical problems, of quickly growing groups of electro-hypersensitives, mainly youngsters.

						That is exactly why people don't trust AI. What solution? The High-Level Expert Group should initiate European laws to enshrine and protect trustworthiness of AI on a legal basis, a legal fixation, the offence against which could be prosecuted at the courts (follow-up of DSGVO of May 2018).	In its second working-period the High-Level Expert Group should take on board these groups: - to formulate the legal basis of trustworthy AI, made in Europe - to tackle the immense problems of millions of job-losers by introducing AI-Systems with unpredictable troubles for societies, which cannot be met by European tax payers.
Tomislav	Bracanovic	Institute of Philosophy / Faculty of Electrical Engineering and Computing, Zagreb (external associate) / COMEST-UNESCO (member)	[1] On p. i, it is claimed that "trustworthy AI will be our north star" and, on p. 1, that "we therefore set Trustworthy AI as our north star". The "north star" metaphor – as metaphor for some sort of a beacon providing guidance or orientation – might be inappropriate in this context. It seems to imply that, in the future, AI itself is supposed to shape the way we humans are going to deal with such technologies. Although "north star" metaphor (when applied to AI) may be just a matter of style, it leaves the impression of inconsistency with strong (and justified) emphasis that this document places on human "values", "autonomy" and "oversight". If the metaphor is retained in the final version of the document, it should at least be clarified whether "trustworthy AI will be our north star" (as it is claimed in the introduction) or it "has been our north star" (as it is claimed in the conclusion). [2] The phrase "human-centric" is correctly explained (in Glossary, p. iv) and is contextually clear. Nevertheless, since "human-centric" means practically the same as "anthropocentric", and "anthropocentrism" (as opposed to "biocentrism") is a technical term strongly associated with one particular position in environmental ethics, it might be prudent to use the more neutral term "human-centered" (among other things, to avoid the impression that the authors of this document are somehow taking side in this longstanding environmental ethics debate).	[1] On p. 5, it is claimed that AI HLEG "believes in an approach to AI ethics that uses the fundamental rights commitment of the EU Treaties and Charter of Fundamental Rights as the stepping stone to identify abstract ethical principles…" Also, on p. 6, it is claimed that "fundamental rights provide the bedrock for the formulation of ethical principles" and that "AI HLEG is not the first to use fundamental rights to derive ethical principles and values". However, on p. 5, it is claimed that "ethics is the foundation for, as well as a complement to, fundamental rights endorsed by humans". This appears contradictory, because it suggests both (a) that ethics (and its principles) lies at the foundation of fundamental rights and (b) that fundamental rights are the bedrock (foundation) of ethical principles (or the source from which ethical principles are derived). The exact nature and mutual relationship between "rights" and "moral principles" or "values" is a perennial philosophical problem and rephrasing the relevant sentences might be advisable (perhaps emphasizing only the complementarity of fundamental rights and ethical principles, without going into the question which of them is more fundamental). [2] On p. 5, it is claimed: "The field of ethics is also aimed at protecting individual rights and freedoms, while maximizing wellbeing and the common good." Suggestion: change the beginning of this sentence so as to read something like: "Ethical endeavors are also aimed at protecting individual rights and freedoms, while maximizing wellbeing and the common good." Speaking about ethics as a "field" suggests (quite correctly) that ethics is an academic field (i.e. branch of philosophy). Academic fields, however, primarily aim to achieve certain theoretical truths and many philosophers (academic ethicists included) would reject interpreting their work as any sort of activism (no matter how justified it may be). And again, the sentence should be changed also because within the "field of ethics" there are so many diverging views about "rights", "freedom" and "well-being". [3] On p 7, it is claimed: "To specify the development or application of AI in line with human dignity, one can further articulate that AI systems are developed in a manner which serves and protects humans' physical and moral integrity, personal and cultural sense of identity as well as the satisfaction of their essential needs." A suggestion: consider adding a caveat into this sentence, after the word "identity", perhaps something like "(as long as it does not harm others or infringes on other's rights)" [4] On p. 9, it is claimed: "Of equal importance, AI systems should be developed	[1] On p. 20, it is claimed: "An intelligent system that will have the capabilities to learn and adapt its behaviour actively can be understood as a stochastic system and is often described by a 'sense-plan-act' cycle. For such architecture to be adapted to ensure Trustworthy AI, ethical goals and requirements should be integrated at 'sense'- level in a way that plans can be formulated that observe and ensure adherence to those principles. In this way, actions and decisions by the system reflect the observed principles." Why should "ethical goals and requirements" be integrated only at "sense" level? Why not at "plan" and, especially, "act" level too? Some concrete example might help to clarify this. [2] On p. 20, it is claimed: "While traceability is not (always) able to tell us why a certain decision was reached, it can tell us how it came about – this enables reasoning as to why an AI-decision was erroneous and can help prevention of future mistakes." Two things are unclear with this sentence: (a) It sounds odd to claim that "traceability" (an abstract noun) is able or unable to "tell us" anyting. (b) I do not see the difference between "why a certain decision was reached" and "how [this decision] came about". [3] On p. 22, it is claimed: "We invite stakeholders partaking in the consultation of the Draft Guidelines to share their thoughts on additional technical or non-technical methods that can be considered in order to address the requirements of Trustworthy AI." I believe this document did a good job in covering such methods. The only thing that comes to my mind at the moment: perhaps introducing, at the EU level, a kind of a label or certificate (an analogue to Michelin stars?) indicating that the particular AI product is designed in accordance with Trustworthy AI principles or values.	No comments here.	The document, as I believe, is well-structured and intended, especially as it attempts to shift focus from general considerations (abstract ethical values, principles, rights etc.) to specific issues of their application in concrete cases in the real world. I also find commendable its balanced approach to both potentially good and potentially bad uses of AI based technologies. Dr. Tomislav Bracanović Research Associate, Institute of Philosophy, Zagreb External Associate (course "Ethics and New Technologies"), Faculty of Electrical Engineering and Computing, University of Zagreb Member and rapporteur of COMEST (UNESCO's World Commision on the Ethics of Scientific Knowledge and Technology)

and implemented in a way that protects societies from ideological polarization and algorithmic determinism." It is unclear what is meant by "algorithmic determinism". And more importantly: to propose developing and implementing AI systems so as to "protect societies from ideological polarization" may sound like a proposal to control or limit people's freedom of thought and expression. Since ideological differences and disagreements are part and parcel of democratic society, I would suggest to change (maybe even omit) this sentence.

[5] On p. 9, it is claimed: "Therefore not only should AI be designed with the impact on various vulnerable demographics in mind but the above mentioned demographics should have a place in the design process (rather through testing, validating, or other)." It is unclear in which way should (actually could) the mentioned demographics (especially children and immigrants) have place in the process of design of AI technologies. The final part of the sentence ["(rather through testing, validating, or other)"] is a bit unclear.

[6] On p. 10, it is claimed: "Furthermore, to ensure human agency, systems should be in place to ensure responsibility and accountability. It is paramount that AI does not undermine the necessity for human responsibility to ensure the protection of fundamental rights." The first sentence, as I believe, should be revised because it suggests that "responsibility" and "accountability" are preconditions for (come before) "human agency". However, it is the other way around: "responsibility" and "accountability" are possible (they make sense) only after human agency is ensured (as the capacity to act on the basis one's free and rational choice). The second sentence is somewhat convoluted ("…does not undermine the necessity for human responsibility to ensure…") and perhaps should be reworded.

[7] On p. 12, it is claimed: "Normative citizen scoring (e.g., general assessment of 'moral personality' or 'ethical integrity') in all aspects and on a large scale by public authorities endangers these values, especially when used not in accordance with fundamental rights, or when used disproportionately and without a delineated and communicated legitimate purpose. Today, citizen scoring – at large or smaller scale – is already often used in purely descriptive and domain-specific scorings (e.g. school systems, e-learning, or driver licenses)." The mention of "citizen scoring" in the first sentence has this strong (Big Brother-like) negative connotation – which is surely justified as long as such "citizen scoring" would indeed assess one's moral personality or, e.g., loyalty to the state. It seems unjustified, however, to transfer this highly negative connotation of "citizen scoring" (by using the same expression in the second sentence), to assesments made in school systems or when issuing driver's licenses.

Anonymous	Anonymous	Anonymous	Wenn die Europäische Kommission zu ihrem Entwurf ihrer Ethik-Leitlinien für Künstliche Intelligenz zur Stellungnahme aufruft, wie in ihrer Pressemitteilung vom 12.12.18 und erneut am 15.1.2019 geschehen, ist nicht erklärlich, weshalb die konkrete Konsultation sich auf den Standpunkt stellt, dass es sich bei den Richtlinien (!) nicht um Standpunkte der Europäischen Kommission, sondern eines Expertengremiums handele. Hier erwarte ich als EU-Bürgerin keine Flucht ins Unverbindliche, sondern Übernahme der Verantwortung für – notwendige – Richtlinien zum ethischen Umgang mit künstlicher Intelligenz. „The Guidelines are addressed to all relevant stakeholders developing, deploying or using AI, encompassing companies, organisations, researchers, public services, institutions, individuals or other entities. In the final version of these Guidelines, a mechanism will be put forward to allow stakeholders to voluntarily endorse them." „A mechanism will be put in place that enables all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis. This will be set out in the final version of the document." Welchen Sinn hat bei der Entwicklung ethischer Prinzipien die Freiwilligkeit deren Beachtung? In einer Realität, die durch ausgeprägte Asymmetrie der Einflussmöglichkeiten gekennzeichnet ist – Verbraucher auf der einen Seite mit praktisch kaum einer Wahl – auf der anderen Seite große Unternehmen, deren Algorithmen als Geschäftsgeheimnisse geschützt werden (vgl. die Rechtsprechung des BGH zur Schufa, die die Algorhythmen zu schützenswerten Geschäftsgeheimnissen erklärte) halte ich das Abstellen auf Freiwilligkeit für reine Alibitätigkeit. Ich halte es für höchst unwahrscheinlich, dass durch Einhalten der Richtlinien Verbraucher in die Lage versetzt würden, die Tragweite ihrer Handlung bei einer Einwilligung zu ermessen. Schließlich fehlt es den Verbrauchern auch an Möglichkeiten zur Kontrolle. Das entspricht – anders als auf S. 2 dargestellt – gerade keiner wirklichen effektiven Willensausübung.	Auf S. 7 heißt es: „Citizens should never be subject to systematic scoring by government." Dieser Aussage stimme ich in vollem Umfang zu, vor allem, wenn ich an die Entwicklungen in China denke, wo intensiv an einem Bürger-Scoring gearbeitet wird. Im gleichen Atemzug frage ich mich, weshalb diese Einschränkung nur für Regierungen gelten soll. Aus meiner Überzeugung darf es auch Unternehmen nicht erlaubt werden, den Bürger gläsern zu machen, in dem Scoring dazu führt, dass alle verfügbaren Daten – ohne dass der Einzelne die realistische Möglichkeit hätte, deren Richtigkeit zu prüfen – für Kreditprüfungszwecke, Entscheidungen über Arbeitsverhältnisse, Versicherungen u.ä. verwendet werden. In die gleiche Richtung geht die Argumentation auf S. 17, wenn es heißt: „Transparency is key to building and maintaining citizen's trust in the developers of AI systems and AI systems themselves. Both technological and business model transparency matter from an ethical standpoint. Technological transparency implies that AI systems be auditable,14 comprehensible and intelligible by human beings at varying levels of comprehension and expertise. Business model transparency means that human beings are knowingly informed of the intention of developers and technology implementers of AI systems." Auch hier reicht es nicht, auf die Prüffähigkeit von Algorhythmen abzustellen, wenn nicht gleichzeitig sichergestellt wird, dass es unabhängige Prüfinstanzen gibt, die Algorhythmen auch prüfen können (und es nicht wie im bereits zitierten BGH-Urteil unter Verweis auf Geschäftsgeheimnisse verwehrt wird). Hier wäre wenigstens eine Kontrolle eine Art „TÜV" von Nöten, um ethische Prinzipien bei der Verwendung der Algorhythmen zu ermöglichen. Verbraucher könnten dann auch ihre Produkt/Dienstleistungsentscheidung an Hand des „TÜV-Siegels treffen. Ohne neutrale Kontrolle bleibt es bei der außerordentlich asymmetrischen Rolle des Verbrauchers /Bürgers einerseits und der Stelle, die hinsichtlich ihrer Algorhythmen keine Kontrolle befürchten muss.	Ins Leere werden auch Verhaltensregeln laufen, wie auf S. 22 „Codes of Conduct" formuliert. Als Verbraucherin habe ich keine Möglichkeit, zu kontrollieren, was z.B. bei den Regeln des autonomen Fahrens programmiert wurde, wenn die Entscheidung zwischen der Gefährdung von Insassen und von Passanten zu treffen ist.	Wenn die Europäische Kommission eine breite Partizipation ihrer Bürgerinnen und Bürger wünscht, hätte sie die gleichen Schritte einleiten können wie bei der Beteiligung der Bevölkerung an der Meinungsbildung zur Abschaffung der Sommerzeit. Zur Erinnerung: 5 Millionen Menschen haben daran teilgenommen. Wenn es der Europäischen Kommission nicht gelingt ein Papier, das mehr als 20 Seiten umfasst, aber ein so wichtiges Thema wie den ethischen Umgang mit künstlicher Intelligenz umfasst, nicht in den Amtssprachen der EU veröffentlicht, lässt dies nur den Schluss zu, dass eine gründliche Befassung der Bevölkerung nicht erwünscht ist. Dieser Eindruck drängt sich auch bei der Lektüre derjenigen auf, die in dem von der Kommission berufenen Gremium sitzen und den Entwurf der Richtlinien verfasst haben. Was haben IBM und Google als US-amerikanische Großkonzerne mit der ethischen Verantwortung innerhalb der europäischen Union zu tun? Warum fehlen in dem Gremium kritische Fachleute wie z.B. aus dem Chaos-Computer-Club? Ebenso wenig überzeugt der Zeitpunkt der Öffentlichkeitskampagne, die zunächst auf einen Monat begrenzt wurde und am 18.12.18, also kurz vor den Weihnachtsfeiertagen begonnen wurde und den Schluss nahe legt, dass man nur eine pro-forma Beteiligung macht. An diesem Eindruck ändert auch die minimale Verlängerung der Beteiligungsfrist nichts. Vielmehr erweckt der erzeugte Zeitdruck unangenehme Erinnerungen an das Verfahren im Trilog-Verfahren bei der Datenschutzgrundverordnung. Auch dort wurde ein künstlicher Zeitdruck aufgebaut, der im Ergebnis dazu geführt hat, dass die DSGVO zahlreiche handwerkliche Mängel enthält und die DSGVO an der Marktasymmetrie, die im Internet herrscht, gerade in Bezug auf die Großen der Branche wie Google, Facebook, Amazon nicht wirklich etwas geändert hat. Heutzutage können Sie nicht einmal mehr ein Antivirusprogramm kaufen, ohne dass der Anbieter von Ihnen verlangt, dass Sie Ihre Daten zu Tracking-Zwecken zur Verfügung stellen. Als Verbraucher haben Sie keine wirkliche Wahl, weil es eben keine Antivirenprogramme gibt, die auf Tracking verzichten. Art 7 der DSGVO läuft insoweit völlig leer. Es heißt zwar in den Vorbemerkungen zum Ethik-Richtlinienentwurf als Ziel der Richtlinien: „Provide, in a clear and proactive manner, information to stakeholders (customers, employees, etc.) about the AI system's capabilities and limitations, allowing them to set realistic expectations" Wenn das tatsächlich das Ziel ist, bedarf es einer erheblich größeren Beteiligung der Bevölkerung, um für Informationen und Diskussionen zu sorgen.

Anonymous	Anonymous	Anonymous	1. First of all, thank you for being pro-active regarding robotics & AI and leading the way. However, these ethics guidelines can only have effect and promote trustworthy innovation, if the EU also starts fixing the current issues we already face with tech giants that dominate the market and have (shown to have) the power to place themselves above any ethical principle. 2. How does this document (Ethics Guidelines for Trustworthy AI) relate to the Resolution on Civil Law Rules of Robotics (2015/2103(INL)?3. The industry has welcomed several ethic codes regarding AI , however, since AI is a global issue and to a very large extent, transboundary in its reach and approach, shouldn't all these (different) ethical principles regarding AI (including the new HLEG guidelines), be geared to one another as much as possible? This is also important in view of the fact that one commits itself to these ethic codes by signing.	4. Page 10, 1st paragraph: "If one is a consumer or user of an AI system this entails a right to decide to be subject to direct or indirect AI decision making, a right to knowledge of direct or indirect interaction with AI systems, a right to opt out and a right of withdrawal." + "opt out" in general.Comment: The opt-out option is important, however, may have as result that the AI-based service will be denied to the consumer/user, without having any alternatives (in the AI-driven world), which could lead to social exclusion. Then effectively there is no "opt out".5. Critical concerns raised by AI page 11:Here I would like to suggest two cases as regard to particular uses or applications, sectors or contexts of AI that may raise specific concerns: (A) Loomis v. Wisconsin, 137 S. Ct. 2290 (2017). -> The ethical arguments for and against algorithms in sentencing. Reasons for concern about the transparency, accuracy, fairness and accountability of COMPAS. + algorithmic due process.(B) Kosinski, M, & Wang, Y. (2017) Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images, Journal of Personality and Social Psychology. February 2018, Vol. 114, Issue 2, Pages 246-257. The research shows that Deep Neural Networks are very accurate at detecting sexual orientation from facial images. Although this kind of research may be harmless on itself (face recognition) and perhaps fall in line with HLEG Guidelines, it is very likely going to be used for less honorable purposes. Or, as the researchers wrote on the side-effects of their findings, "[..] given that companies and governments are increasingly using computer vision algorithms to detect people's intimate traits, our findings expose a threat to the privacy and safety of gay men and women." 6. 5.4, page 12 LAWS "Note that, on the other hand, in an armed conflict LAWS can reduce collateral damage, e.g. saving selectively children." Comment: People would probably exploit this knowledge by using children as a human shield, just like we saw in the Syria conflict.	7. Page 16, "6. Respect for (& Enhancement of) Human Autonomy"An other serious concern here is the filter bubble being created by those agents which causes the average user never to see results or possibilities outside his/her comfort zone.	8. Page 24, (6. Respect for Privacy): "If applicable, is the system GDPR compliant?" The question is if this is always really possible. In the book 'Robot-is-me?' (2018), chapter 7, the author explains why the GDPR may not be sufficient to address problems that will arise from using autonomous AI on the IoT(which also eliminates the need for human intervention and/or the need for human-in-the-loop). For example responsibility and accountability could, under circumstances, be escaped from somewhere in the controller-processor chain. This comment also applies to page 17, item 7.	9. There's one topic I would like to suggest that has not been covered in this document yet and that is the development of in-body AI-devices that directly connect with human brains. Now more and more companies are focusing on this type of AI-applications that can influence our neurons in a specific portion of the brain and given the intrusive nature of the application, it is perhaps not an unnecessary exercise to formulate additional ethical requirements for this.
Anonymous	Anonymous	Anonymous					

Giacomo	Tesio	The first lines of the introduction highlight a serious flaw of the draft: the pillars that underpin the Commission's vision show a fundamental bias: 1. increasing public and private investments in AI to boost its uptake, 2. preparing for socio-economic changes, and 3. ensuring an appropriate ethical and legal framework to strengthen European values Being eager to adopt a largely misunderstood technology obviously inhibits the ability to reason about its limits and risks. Before trying to boost its uptake, the Commission should try to understand to what extent and in which fields of endeavour the set of techniques that goes under the AI umbrella should be experimented. As Shoshana Zuboff recently wrote, technology is NOT an unstoppable force of nature, but a human artifact serving interests and needs of specific humans. In other words, Technology is a prosecution of Politics by other means: each advancement can be designed to serve the public interest or private and elitarian ones. And just like with Politics, a renounce to participate to its course just means to being subject to others' will. Before talking about "Trustworthy AI", we should have a population able to understand the topic enough for their trust to be meaningful. As for today, without a serious investments in schools to foster History and Informatics as preconditions of our citizenship, such trust can not be meaningful but just deceptive and ill founded. It's not a trust on the technology, but in the corporations and the "experts" that can exploit such trust and the widespread ignorance of the topic to weaken regulations and streghten their handle on society. Having said that, the high level description outlined for the Ethical framework is basically sound: it's reasonable to think that when the whole population will be able to understand how a neural network's calibration differs from a k-mean clustering, a similar framework will emerge. However the glossary that preceed the Introduction already shows that we are not ready for such framework: despite being written by an high level expert group on AI, the definitions still use an antropomorphic language to describe what is just software. In particular describing software bugs (either intentional or unintentional) as "bias" shows a deep misunderstanding about the software in question and about the statistical processes that define its behaviour.	Despite an interesting and convidisible introduction, the principles that the chapter proposes lack a fundamental hierarchical structure. It should be quite evident by looking at such principles: - The Principle of Beneficence: "Do Good" - The Principle of Non maleficence: "Do no Harm" - The Principle of Autonomy: "Preserve Human Agency" - The Principle of Justice: "Be Fair" - The Principle of Explicability: "Operate transparently" Even if we hadn't more than two thousands years from the Hippocratic Oath and generations of physicians grown with the "Primum non nocere" maxim, we can see how the last three principles are just specializations of the more general "Do no Harm". In particular the Principle of Autonomy tries to address risks to individuals, the Principle of Justice tries to address the risks to weak groups and the Principle of Explicability tries to address socio-political risks. Since the Principle of Non Maleficience is so preponderant to require three specializations, we should put it first, before the principle of Beneficience, and underlining its relation with the others: - The Principle of Non maleficence: "Do no Harm" - The Principle of Autonomy: "Preserve Human Agency" - The Principle of Justice: "Be Fair" - The Principle of Explicability: "Operate transparently" - The Principle of Beneficence: "Do Good" The road to hell is paved with good intentions: just like with medicine, whenever simpler and safer solutions exist they should be preferred to more complex and risky ones. But there is an even more important omission in the list: the Principle of Ultimate Human Accountability. This is a fundamental principle that underlie all European ethical and legal system: at least a human must always be accountable for the problems caused by a human artifact. In other terms: what is forbidden to a human can not be allowed through an artificial proxy, no matter how "autonomous" (aka expensive to debug) such proxy is. Talking about ethics is void if we are not ready to enforce this simple but fundamental principle of human responsibility. The section on "Lethal Autonomous Weapon	Even this chapter present several issues: 1. The short paragraph about "Accountability" suggest to design mechanisms that can range from monetary compensation to apology, but it forgets to include prison: to gain trust it is important to explicitly state that an autonomous proxy cannot become a "Get Out of Jail Free" ticket. 2. The section on "Safety" looks like it was designed to be ineffective: it's pointless to assess potential risks associated with the use of AI-based products and services without defining serious punishments when things goes wrong anyway. 3. The section on "Trasparency" is too vague and forgiving: a simpler approach is to say that no opacity must be allowed in applications that consume human data. Such rules would instantly skyrocket private and public investments in AI research, looking for new machine learning techniques that can be fully explained and debugged. 4. The section on "Robustness" looks well designed but open to a wide de-responsibilization when it improperly talks about "non-determinism" (false, if we are talking about deterministic, non-quantum, computers) and it cites "complexity", "opacity", and "sensitivity to training/model building conditions" as a sort of justifications for unreproducible results. Simply, whenever such conditions exists, the AI program is not robust and should not be applied to problems that require such robustness. 5. The section about "Human Autonomy" is very scary: in no way people should be nudged by machines. If AI will be successful in enhancing human wealth as it's promise to be, a lot of friendly people with a lot of free time will be able to nudge us on our request, but it's too dangerous to let flawed machines manipulate humans whatever the goal: every software has bugs vulnerabilities and many have intentional backdoors: AI won't be different. Later, in "Architectures for Trustworthy AI", while considering the technical means to ensure an ethical behaviour the HLEG suggest to integrate an ethical signal in the "sense" phase of the stochastic system. This is both naive and weird: - WHICH ethics we should use? If we widely deploy autonomous machines following a certain ethical model, people will adapt to it (because machines cannot really adapt to us): this could turn to be most effective brain washing project ever conceived. Humans naturally adapt to the sorrounding intelligences: put a consumist agent in every room, and you will build a population of consumists. - HOW MUCH ethics? Who will decide the weight of that signal? And what when a bug will inhibit it?	I really appreciated the flexible approach to the assessment process: talking about ethics, a checklist would be too easy to exploit. For sure, each technique requires different kind of assessments: for example the dataset used to calibrate a k-mean could be enough to reproduce the calibration process and to exclude any racial discriminations, but it would be totally inadequate for assess any property of a classifier based on an artificial neural network. The risk however is that, without a widespread understanding of the AI techniques, the Commission will ask to the wolfs how to rule the sheeps: we cannot rely on experts that consults large corporations to define any assessment of "trust" into something that can manipulate people. Moreover, being able to assess the Ethics of a "Trustworthy AI" cannot replace clear regulation establishing the characteristics that an algorithm must have before being fed with human data. In particular we need to extend the right to "meaningful information about the logic involved" by each AI processing beyond the individuals protected by the article 13 of the GDPR: even groups, such as families, neightbors, customers and so on should have the right to know and understand the exact logic applied to their collective data, when and to which aim the processing occurs.	Despite all the issues described above, I appreciated the effort and care that has been evidently put by the HLEG in the writing of this draft. It's important for Europe to fill our technological gap with U.S.A. and China and it's conforting to see serious people working on the ethical issues that will emerge from the AI adoption. However is even more important to avoid short-cuts. Good will and honesty are fundamental, but not enough to balance lobbying and hype. To address our technological issues (including AI adoption) we need to raise the general population understanding of Informatics. We need a new mass education plan, with serious investments on teachers and professors from the primary school on. We need to raise a generation of people able to modify the software that they use and they feed with their own data. Since Technology is Politics, being able to self-host and customize the applications we use is the only way to preserve democracy: it will prevent data capitalization and people manipulation. Programming is today what Writing was during Ancient Egypt: a tool which is totally primitive, but effective to collect and retain Power among humans exactly because it is primitive. We need better systems, better programming languages and people able to use software without being manipulated through it. Until then, widespread adoption of AI can be useful, but it's irresponsible to apply it to human data. We need prudent regulations that err on the side of caution, not because computer-aided statistics is dangerous in itself but because it's too easy to abuse it and manipulate or hurt people and societies in a context when most people can't understand their working.

			Later on, similar concerns emerge when the draft cites "non-determinism" while talking about software that is executed by deterministic machines (aka computers). Such language is worrying because it shows a tendency from the HLEG to rationalize the risks as inevitable instead of understanding them deeply and taking them into account.	Systems" is in direct contrast to all the principles stated above. The only way an Ethical Framework can be credible while proposing principles like "Do no Harm", "Preserve Human Agency", "Be Fair", "Operate transparently" and "Do Good" is to clearly state that Autonomous Weapon Systems (lethal or not) must be forbidden on the European territory. The section on the "Potential longer-term concerns" shows the usual sci-fi based fears that are the flip side of the current hype. Instead of being concerned about Artificial Consciousness that would be way easier to fake than to implement we should be afraid of semi-autonomous weapons in the hands of a small group of people holding most of the planet's wealth. And in the count of such weapons we should obviously include every tool that can be used to direct human attention, to manipulate feelings or perceptions and to forge mass opinions.	Or what if other inputs overwhelm such signal? The only use of an ethical signal in an autonomous system is to shield corporations from taking full responsibility of errors: it's dumb to pretend to teach ethics to trolleys, we should build infrastructures that simply prevent lethal incidents to occur. Moreover in the section about Regulation we lack any reference to penal justice: just like before, it should be clearly stated that when an autonomous artifact kill or harm, one or more humans will be held fully accountable for it.	
Janis	Ratkevics	Ministry of Environmental Protection and Regional Development of Republic of Latvia					Ethical guidelines should include the point that an AI system should always alert a person if person activities can cause harm to himself, other people or living environment.Highlight that the point on the restrictions of autonomous lethal systems will work only if it is accepted by all the big players. Otherwise, the point will be declarative.
Carmen	Mac Williams	Grassroots Arts and REsearch	On page 2 in the 3rd line it is stated that "…it should be noted that no legal vacuum currently exists, as Europe already has regulation in place that applies to AI…" My Comment: Please give a reference to these existing regulations. In my knowledege there is now a first German regulation demaded by the German Ethical Commission (see: https://www.bmvi.de/SharedDocs/DE/Pressemitteilungen/2017/128-dobrindt-massnahmenplan-ethikregeln-fahrcomputer.html) on automatic driving, but what are the other regulations. I do not think we yet have globally or in Europe many legal regulations on AI. So please specify it in your report, where this professional Training of external ethical experts shall take part and how a AI Company can find these external ethical experts. I can only refer to Kate Crawford, co-author of the must-read 'AI Now 2017 Report.:		§ Traceability & Auditability Page 20: My comment concerns following Text in the draft "Evaluation by internal and external auditors can contribute to the laymen's acceptance of the technology. Importantly, in order to enable regulatory bodies to undertake verification and auditing of AI systems where needed,.." My Comment: There has to be a gigantic effort to Train EXTERNAL Ethic AI advisors as this Profession does not exist. Please clarify how this Training of External Advisors should be achieved . It is not the answer to say that typical global Consulting firms will do this as they only do what the customer wishes. So they hardly arer guided by human centric, good for Society, ethical guidelines. So who will be These ethical advisors externally?	Concerning to the text in the draft on page 28: "We invite stakeholders partaking in the consultation of the Draft Guidelines to share their thoughts on additional technical or non-technical methods that can be considered in order to address the requirements of Trustworthy AI." My comment: Trustworthy AI requires understandable explanation by the technical expert of the AI system to the non-technical member of society (e.g. users, customers, trainers, managers, employees, politicians), what the system does and for what it us usable. It is not enough that the AI system explains to the technical skilled individual what it does, but it is also important that the technical skilled individual can explain it to the non-technical skilled individual. So again we Need Training for AI uman technical experts to explain to non-technical People, what the AI System does. RThis is communication skills which most technical People lack. So where shall this professional	The key is to create binding ethical regulations for Europe (ideally for the world), which can be legally enforced. The external ethical advisors is qualified to issue a Q/A Branding of "Trustworthy AI", which adds value to the AI system, as the potential customer can choose this AI System instead of another unlabelled, therefore not to be trusted, system.

		"What is most urgently needed now is that these ethical guidelines are accompanied by very strong accountability mechanisms. We can say we want AI systems to be guided with the highest ethical principles, but we have to make sure that there is something at stake. Often when we talk about ethics, we forget to talk about power".		Training taking part?	
Joachim	Urbanek				Stellungnahme 1. Anmerkungen zum Verfahren Wenn die Europäische Kommission eine breite Partizipation ihrer Bürgerinnen und Bürger wünscht, hätte sie die gleichen Schritte einleiten können wie bei der Beteiligung der Bevölkerung an der Meinungsbildung zur Abschaffung der Sommerzeit. Zur Erinnerung: 5 Millionen Menschen haben daran teilgenommen. Wenn es der Europäischen Kommission nicht gelingt ein Papier, das mehr als 20 Seiten umfasst, aber ein so wichtiges Thema wie den ethischen Umgang mit künstlicher Intelligenz umfasst, nicht in den Amtssprachen der EU veröffentlicht, lässt dies nur den Schluss zu, dass eine gründliche Befassung der Bevölkerung nicht erwünscht ist. Dieser Eindruck drängt sich auch bei der Lektüre derjenigen auf, die in dem von der Kommission berufenen Gremium sitzen und den Entwurf der Richtlinien verfasst haben. Was haben IBM und Google als US-amerikanische Großkonzerne mit der ethischen Verantwortung innerhalb der europäischen Union zu tun? Warum fehlen in dem Gremium kritische Fachleute wie z.B. aus dem Chaos-Computer-Club? Ebenso wenig überzeugt der Zeitpunkt der Öffentlichkeitskampagne, die zunächst auf einen Monat begrenzt wurde und am 18.12.18, also kurz vor den Weihnachtsfeiertagen begonnen wurde und den Schluss nahe legt, dass man nur eine pro-forma Beteiligung macht. An diesem Eindruck ändert auch die minimale Verlängerung der Beteiligungsfrist nichts. Vielmehr erweckt der erzeugte Zeitdruck unangenehme Erinnerungen an das Verfahren im Trilog-Verfahren bei der Datenschutzgrundverordnung. Auch dort wurde ein künstlicher Zeitdruck aufgebaut, der im Ergebnis dazu geführt hat, dass die DSGVO zahlreiche handwerkliche Mängel enthält und die DSGVO an der Marktasymmetrie, die im Internet herrscht, gerade in Bezug auf die Großen der Branche wie Google, Facebook, Amazon nicht wirklich etwas geändert hat. Heutzutage können Sie nicht einmal mehr ein Antivirusprogramm kaufen, ohne dass der Anbieter von Ihnen verlangt, dass Sie Ihre Daten zu Tracking-Zwecken zur Verfügung stellen. Als Verbraucher haben Sie keine wirkliche Wahl, weil es eben keine Antivirenprogramme gibt, die auf Tracking verzichten. Art 7 der DSGVO läuft insoweit völlig leer. Es heißt zwar in den Vorbemerkungen zum Ethik-Richtlinienentwurf als Ziel der Richtlinien: „Provide, in a clear and proactive manner, information to stakeholders (customers, employees, etc.) about the AI system's capabilities and limitations, allowing them to set realistic expectations"

Wenn das tatsächlich das Ziel ist, bedarf es einer erheblich größeren Beteiligung der Bevölkerung, um für Informationen und Diskussionen zu sorgen.

2. Zur Vorbemerkung
Wenn die Europäische Kommission zu ihrem Entwurf ihrer Ethik-Leitlinien für Künstliche Intelligenz zur Stellungnahme aufruft, wie in ihrer Pressemitteilung vom 12.12.18 und erneut am 15.1.2019 geschehen, ist nicht erklärlich, weshalb die konkrete Konsultation sich auf den Standpunkt stellt, dass es sich bei den Richtlinien (!) nicht um Standpunkte der Europäischen Kommission, sondern eines Expertengremiums handele. Hier erwarte ich als EU-Bürgerin keine Flucht ins Unverbindliche, sondern Übernahme der Verantwortung für – notwendige – Richtlinien zum ethischen Umgang mit künstlicher Intelligenz.

„The Guidelines are addressed to all relevant stakeholders developing, deploying or using AI, encompassing companies, organisations, researchers, public services, institutions, individuals or other entities. In the final version of these Guidelines, a mechanism will be put forward to allow stakeholders to voluntarily endorse them." „A mechanism will be put in place that enables all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis. This will be set out in the final version of the document."
Welchen Sinn hat bei der Entwicklung ethischer Prinzipien die Freiwilligkeit deren Beachtung? In einer Realität, die durch ausgeprägte Asymmetrie der Einflussmöglichkeiten gekennzeichnet ist – Verbraucher auf der einen Seite mit praktisch kaum einer Wahl – auf der anderen Seite große Unternehmen, deren Algorithmen als Geschäftsgeheimnisse geschützt werden (vgl. die Rechtsprechung des BGH zur Schufa, die die Algorhythmen zu schützenswerten Geschäftsgeheimnissen erklärte) halte ich das Abstellen auf Freiwilligkeit für reine Alibitätigkeit. Ich halte es für höchst unwahrscheinlich, dass durch Einhalten der Richtlinien Verbraucher in die Lage versetzt würden, die Tragweite ihrer Handlung bei einer Einwilligung zu ermessen. Schließlich fehlt es den Verbrauchern auch an Möglichkeiten zur Kontrolle. Das entspricht – anders als auf S. 2 dargestellt – gerade keiner wirklichen effektiven Willensausübung.
3. Zu den Richtlinien
Auf S. 7 heißt es:
„Citizens should never be subject to systematic scoring by government."
Dieser Aussage stimme ich in vollem Umfang zu, vor allem, wenn ich an die Entwicklungen in China denke, wo intensiv an einem Bürger-Scoring gearbeitet wird. Im gleichen Atemzug frage ich mich, weshalb diese Einschränkung nur für Regierungen gelten soll. Aus meiner Überzeugung darf es auch Unternehmen nicht erlaubt werden, den Bürger gläsern zu machen, in dem Scoring dazu führt, dass alle verfügbaren Daten – ohne dass der Einzelne die realistische Möglichkeit hätte, deren Richtigkeit zu prüfen – für Kreditprüfungszwecke, Entscheidungen über Arbeitsverhältnisse, Versicherungen u.ä. verwendet werden.
In die gleiche Richtung geht die Argumentation auf S. 17, wenn es heißt:
„Transparency is key to building and

						maintaining citizen's trust in the developers of AI systems and AI systems themselves. Both technological and business model transparency matter from an ethical standpoint. Technological transparency implies that AI systems be auditable,14 comprehensible and intelligible by human beings at varying levels of comprehension and expertise. Business model transparency means that human beings are knowingly informed of the intention of developers and technology implementers of AI systems." Auch hier reicht es nicht, auf die Prüffähigkeit von Algorhythmen abzustellen, wenn nicht gleichzeitig sichergestellt wird, dass es unabhängige Prüfinstanzen gibt, die Algorhythmen auch prüfen können (und es nicht wie im bereits zitierten BGH-Urteil unter Verweis auf Geschäftsgeheimnisse verwehrt wird). Hier wäre wenigstens eine Kontrolle eine Art „TÜV" von Nöten, um ethische Prinzipien bei der Verwendung der Algorhythmen zu ermöglichen. Verbraucher könnten dann auch ihre Produkt/Dienstleistungsentscheidung an Hand des „TÜV-Siegels treffen. Ohne neutrale Kontrolle bleibt es bei der außerordentlich asymmetrischen Rolle des Verbrauchers /Bürgers einerseits und der Stelle, die hinsichtlich ihrer Algorhythmen keine Kontrolle befürchten muss. Ins Leere werden auch Verhaltensregeln laufen, wie auf S. 22 „Codes of Conduct" formuliert. Als Verbraucherin habe ich keine Möglichkeit, zu kontrollieren, was z.B. bei den Regeln des autonomen Fahrens programmiert wurde, wenn die Entscheidung zwischen der Gefährdung von Insassen und von Passanten zu treffen ist.
Anonymous	Anonymous	Anonymous	• We commend the positive tone in the introduction. • Page i – "…do not aim to provide yet another list…" => should maybe reference examples of such lists?• Page i – "…a mechanism will be put forward to allow stakeholders to voluntarily endorse them." => what incentive or motivation do companies have to do this, or to follow any of the recommendations in the paper? Many companies already have their own Ethical AI Frameworks in place; and while not as onerous as GDPR requirements, compliance with the guidelines set out by the EC would potentially require time and investment in order to set up the necessary processes and procedures, ethical panels/boards etc, especially for early stage companies; may be worth addressing what support or incentives will be available to them? • Page ii – "…living document that needs to be regularly updated…" => should be more explicit perhaps? Frequency, by whom, process…etc	• Page 7 – "…right to be informed of any automated treatment of their data by government bodies…" => this might need to be more explicit, given the pervasiveness of "automation" in processing data, it would not be practical to be informed in every instance; however it does of course make sense in the context of the type of examples that are referred to in the subsequent sentences…• Page 8 – Principle of beneficence: "AI systems should be designed and developed to improve individual and collective wellbeing (…) by generating prosperity, value creation and wealth maximization". This sounds very generic. What does this mean specifically? Any company making more money thanks to AI is in accordance with this principle? • Page 10 – Is it feasible/practical/necessary for consumer to always have "…a right to knowledge of direct or indirect interaction with AI systems…"; the point made on page 3 regarding the need for a tailored approach "…given AI's context-specificity" is relevant here. • Page 10 – "Individuals and groups may request evidence of the baseline parameters and instructions given as inputs for AI decision making…" => while we are in agreement with the importance of the principle of explicability and the need to operate transparently, this needs to be considered in the context of company's IP rights and the need to remain competitive; perhaps a public sector ombudsman/impartial body of sorts could facilitate the needs of both individuals and companies.	• There seems to be some overlap between the 10 principles, e.g. transparency & accountability, design for all & non-discrimination, or safety & robustness. • Page 14 - "Accountability" requirement – perhaps too focussed on the consequences, rather than the method of identifying who is responsible, and how this is determined; differentiate between the AI and the use thereof.• Page 16 - Principle of non-discrimination: intentional discrimination is already prohibited, no need to further specify this with regards to AI. Rely on existing non-discrimination rules. As Chapter I also mentions: no legal vacuum exists.• Page 18 - "…aims to reflect the main approaches that are recommended to implement trustworthy AI." – however some of the implementation methods read more as requirements, rather than truly offering guidance on how to practically implement them; for instance, in the case of "Explanation (XAI research)"• Page 19-20 - Technical methods to achieve trustworthy AI are rather generic, and largely overlap with the principles set out at the start of chapter II. • Page 21 - "Regulation" essentially repeats the "Accountability" requirement on page 14, with little guidance on how to actually implement this• Page 21 – "Explanation" - there is room to elaborate on the circumstances in which it is reasonable to demand an explanation and why and when explanations are useful enough to outweigh the costs;	• The assessment list seems broadly relevant at a high level• Examples of how these are applied to specific use-cases expected in the next draft will certainly be valuable in illustrating its practical application• While many earlier stage start-ups might fall short on these, it probably is still a valuable guide to help foster meaningful discussion about their future capabilities and processes • Some however, such as "Design for all", "Resilience to Attack", and "Respect for Privacy", are not specific to AI, and I think would naturally form part of any software security assessment we would conduct; It's important to keep the scope narrowly focussed on AI, and concise, which might help drive adoption.

Mark	Dugdale	Department of Business, Enterprise and Innovation, Ireland	It would be useful to have a baseline index of existing regulation relevant to AI in Europe. In the paragraphs captioned Trustworthy AI at the end of page 1 and opening of page 2 and again at the top of page 3 in the context of the need for developers to comply with existing regulations. It would be useful to provide references so that the extent of these regulations and the areas where it may be necessary to add to them can be appreciated. On page 3, in the section entitled Scope of the Guidelines', reference is made to sensitivities raised or not raised by certain types of system. The types of systems, envisaged by reference to those recommending songs, may not raise apparent issues according to their purpose but can do so according to the use that the data they collect can be put to. It is reported that some music distribution systems claim to be able to profile their users personality/psychology. This may not be therefore a good comparison to use or alternately should specific reference be made to these very pertinent sensibilities.	On page 5, Section2 - From Fundamental rights to Principles and Values and the passage on Informed Consent it would be useful to make reference to the approach that should be taken to the sometimes black and white, and decidedly coercive, approach to privacy settings taken by some application developers of the 'agree to everything or don't use this at all' variety. In the section, The Principle of Autonomy: Preserve Human Agency, on Page 9, it would be helpful if the ethics guidelines were to address the tension which exists where one person's right to withdraw can effectively cause harm to another by preventing access to a particular potential benefit (e.g. exercising a right to privacy with regard to medical records may mean that there is insufficient data to develop a means of improving diagnosis of a medical condition).	On page 14 under Data Governance. It might be helpful in describing 'inevitable biases' to explain that data drawn from particular societal transactions, such as, recruitment, will reflect prevailing practices which themselves are biased, such as gender balance in the workplace. In the same section, in its final paragraph over on page 15 it should be considered whether, in providing for trust for the data gathering process, by ensuring that data will not be used against those that provide it, in fact creates a form of jeopardy for AI system developers akin to priests in the confessional or therapists at the counselling couch. Given these considerations, it would have to be carefully considered how such a form of 'professional privilege' would be formally recognised and implemented. Further on, on the same page, in '3. Design for all', while it is obviously desirable to try to avoid a 'lowest common denominator' type interface it may then be that this will lead to provision of multiple access facilities which would, by default, cause de facto discrimination according to the users range of skills, abilities and intelligence. In section 4 Governance of AI Autonomy (Human Oversight) also on page 15, two issues arise. Firstly, will it be necessary for ethical advisors to be sector specific so as to be able to identify trigger points in the system that would flag the need for human intervention. Secondly does the very fact that a human overseer can decide to deviate from a system specified outcome raise liability issues for the oversight body relating to the reasonable foreseeability of consequences arising from the new system outputs, where the actions of that overseer result in harm. Also, on page 16 in the final paragraph of '5. Non-Discrimination' it might be considered whether it is possible or beneficial to agree a certain set of characteristics which should be given a null weighting in AI systems (particularly those incorporating machine learning) for example, gender, race and social class. This approach might negate the need to engage in the upstream identification of possible bias in a broad spectrum of cases. In '8. Robustness' on page 17 it is stated that algorithms must be able " to adequately cope with erroneous outcomes". It would be useful to know whether this means that they are able only to reveal the error, explain it or refute it. In the same section and that following, that is to say, '9 Safety', in addition to physical harm and integrity it is suggested that psychological, emotional and reputational harm and integrity also be considered.	On page 25, in paragraph 5. Non-discrimination and the third bullet should "in data and algorithms" might better read "in system outcomes resulting from data inputs, operation of algorithms or howsoever caused" In section 7 Respect for (& Enhancement of) Human Autonomy, page 26. It is suggested that, harking back to the point I made earlier with regard to the use to which data can be put in song selecting systems, a point might be added concerning the need to ensure that data obtained as a result of the operation of the system (i.e personality/psychological profiling) is only put to ethically acceptable uses. In particular are there any red lines in terms of acceptable uses which should be spelled out (or will that be a consideration for the next tranche of work by the AI HLEG - the policy recommendations).
Luca	Scarpiello	EPSU	EPSU agrees with the ETUC/ETUI that there is a need for strong, consistent, and enforceable regulation on AI and recall some of the main elements of their response. AI is the most disruptive technology that we had in several decades, workers and citizens are worried about their future and how that will impact their life and jobs and that of	The ethical purpose of the guidelines promotes a unilateral and voluntary industry-action. Moreover, it limits Union action. A unilateral approach that relies only on industry is not the right way to do this. We need both sides to be involved: emplyers and unions. The EC recognises the limits of self-	The selection of the principles looks only to the side of the developers and there is no mention to the principles of precaution, prevention, solidarity, common good nor distributive justice. We need more than code of conducts, declaration of principles and 'private governance' mechanisms, because this is too	What is important is to implement technology through monitoring mechanisms, so values are effectively respected. Being judge and jury at the same time does not work. Ethical responsibility cannot be left to AI developers. There needs to be an external body to follow developments and to

				regulation in the proposal for a "regulation on promoting fairness and transparency for business users of online intermediation services". Here the EC says clearly "Limiting Union action to promoting voluntary industry-action and certain accompanying measures is possible but unlikely to be effective, as this would essentially rely on the industry's own incentives and willingness to change the status quo.." Respecting ethical values and principles is valuable but it is unlikely to be effective, because it will essentially rely on the industry's own incentives and there is no system to monitor oversight or to solve issues when values get in conflict with other values.	important to be entrusted to developers, companies and innovators without a sanction system. It is too important to base it in code of conduct and principles that are not enforceable. Ethical principles are not associated with any sanction system. the relationship risk/reward is so unbalanced, that some actors may decide that it makes financial sense to break or disrespect the principles. The ETUI has followed the way nanotechnologies have been 'regulated' for the past 10 years, and correctly point out the similarities in that process with the AI 'regulatory' process today. We need cannot end up with a toothless Observatory of AI, like has been done with nanomaterials. We need proper legislation that can be updated to take account of AI developments, including the General Data Protection Legislation, Product Safety Directive, directive on Liability for Defective Products, Directive on Safety at work, and Medical Devices Regulation.	attribute liability. "Minimum regulatory standards need to be developed in order to attribute responsibility and liability in cases where the artificial agent has 'learning and teaching' features and is able to exercise unintended outcomes" (Ponce del Castillo, 2017). An effective regulatory framework is ultimately required in order to ensure that artificial agents co-exist harmoniously with humans and that they are specifically designed for, operating according to human values and needs that are themselves dynamic. Regulators will need to figure out how to manage risks and attribute liability, particularly as machines increasingly acquire the ability to learn and take independent decisions. Without a legal framework, transparency and trust will not exist, which will be detrimental to everyone, including industry. (Ponce del Castillo, 2018).
Anonymous	Anonymous	Anonymous		I 5.2 Covert AI systems: The relevance of this topic is exemplified by the ever-increasing use of chat bots (in either written or vocal communication) where it is not always obvious for the user that the communication partner is not a human. Including the obligation to identify such non-human communication partners into a regulation might be helpful.	II 1.1 Accountability: This section should also discuss how to handle accountability in the case of really severe wrong decisions, e.g. such that cause the loss of human life (Health, autonomic driving etc.). In general, the accountability should be at least as strong as the accountability of a human for the same activity. II 1.2 Data Governance: The paper states "When data is gathered from human behavior, it may contain misjudgement, errors and mistakes. In large enough data sets these will be diluted since correct actions usually overrun the errors, yet a trace of thereof remains in the data". However, this might be too optimistic as the Microsoft chat bot (Tay) failure has demonstrated that one cannot rely on self-correction due to large enough data. II 2 Non-Technical Methods All AI systems should come with a clear description of their limits, including the areas they are intended for and those, they are not intended for, as well as description of input data that the system cannot properly cope with (e.g. an animal recognition system that has been trained with data on mammals might not suit well for identifying insects).	The paper is a valuable step forward in working out and making explicit ethical aspects of AI. However, some adjustments might be necessary,

Anonymous	Anonymous	Anonymous	Dear members of the High-Level Expert Group on Artificial Intelligence,We, the members of the International Committee for Robot Arms Control (ICRAC), have read your draft “Ethics Guidelines for Trustworthy AI” with great interest. We would like to congratulate you on your work in pointing out many of the important ethical-technical issues confronting society. Regarding point 5.4 on p. 12, that is, the paragraph on “Lethal Autonomous Weapon Systems (LAWS)”, we would like to submit three brief comments which we hope to be helpful to your ongoing effort.(1) Given that this is a draft ethics guideline, it would be helpful to include a statement about the most prominent ethical argument on LAWS, namely the unacceptable infringement on human dignity contained in delegating a life-and-death-decision to a machine. Note that according to recent Ipsos polling data (https://www.ipsos.com/en-us/news-polls/human-rights-watch-six-in-ten-oppose-autonomous-weapons), 61% of people surveyed across 26 countries are against LAWS, 66% of whom argue that LAWS would mean crossing a fundamental moral line. The Marten’s clause in international law (Article 1[2] of Additional Protocol 1 to the Geneva Conventions) dictates the consideration of “public conscience” in cases “not covered by the law in force”, such as LAWS, and thus lends further relevance to this wide-spread moral revulsion caused by LAWS. In connection to that, we suggest expanding the following sentence “This raises fundamental ethical concerns,…” as follows: “This raises fundamental ethical, legal and strategic concerns,…”. Our rationale is that the examples you present in the next sentence (arms race, control loss, and lack of fail-safe) are primarily legal and/or strategic rather than ethical in nature. We additionally suggest to add “and responsibility” after “human control” in the final clause of that sentence.(2) We suggest removing the following sentence: “Note that, on the other hand, in an armed conflict LAWS can reduce collateral damage, e.g. saving selectively children.” There is no evidence for this claim. While there certainly are specific effects to be expected from a use of LAWS, one example being an increase in operational speed, there is nothing about full weapon autonomy (i.e. the selection and engagement of targets without human intervention) that predetermines who or what is targeted. Even if, for the sake of the argument, we were to assume recognition algorithms performing accurately (something yet to be demonstrated under battlefield conditions), LAWS could selectively target children just as much as selectively save them. More generally, the reduction of collateral damage depends on the application of military force in line with the principle of distinction (Article 48 and 52 of Additional Protocol 1 to the Geneva Conventions (AP1)) and the principle of proportionality (Article 51(5) (b) AP1). The key issue here is not only the precision of the weapon system in relation to a designated target, but the adjudication of whether or not the target is a lawful one. While an increase in a weapon’s precision can facilitate the application of military force against lawful targets whilst keeping collateral damage as minimal as possible, autonomy in this context is not the same as precision and must not be conflated with it. Just as precision does not require full

				autonomy, full autonomy does not necessarily generate an increase in precision. (3) Note that in the EU Parliament's resolution of 12 September 2018, the EP is not only calling for the urgent development of a common EU position. It also "[u]rges the VP/HR, the Member States and the Council to work towards the start of international negotiations on a legally binding instrument prohibiting lethal autonomous weapon systems" and subsequently "[s]tresses, in this light, the fundamental importance of preventing the development and production of any lethal autonomous weapon system lacking human control in critical functions such as target selection and engagement". Since you "support" this resolution, you might want to consider citing not just its call for a common position, but the actual position on LAWS that the EP is advocating.		
Lê Nguyên	Hoang	EPFL, Science4All	This section was very good.	While many of the fundamental ethics principles presented are compelling, it may be a good idea to argue for further research on what principles should be considered. Better yet, it may be relevant to advocate for research on how to select these principles. This is, for instance, what has been proposed in this research paper: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/17052 But further research seems desirable. Also, in the "critical concerns raised by AI", I fear that the large-scale side effects of social network recommandation systems may have been overlooked. Many researchers have strongly suggested that they cause numerous ethical issues, such as algorithmic biases, filter bubbles, political polarisation, addiction, fakenews propagation, anger proliferation, and so on. Finally, regarding the controversial long term issues, rather than relying on individual intuitions, it may be worthwhile to point to surveys of the experts, such as this one: https://arxiv.org/abs/1705.08807 Intriguingly, the median expert assigns a 10% probability to human-level AI by 2025. Caveats definitely apply. But the order of magnitude of this probability makes it far less negligible than, say, death by car accident before 2025.	It may be relevant to just encourage developers to care about user experience and social consequences, especially if the intent of the document is to provide guidelines rather than verifiable rules. In particular, it may be worthwhile to argue that large-scale AIs, especially when deployed on the internet, usually have social consequences that are extremely hard to predict ahead of time. Without much thinking, such consequences will likely be overlooked.	This section is great.	It seems to me that recommendation systems may have been overlooked. Yet, these days, and in the foreseeable future, such AIs may be those that have the greatest social consequences. I believe that more thought should be given to them.
Yiannis	Kanellopoulos	Code4Thought P.C.		I do like the five ethical principles described in this chapter and the fact they're aggregating a set of more values and principles. 5.3 Normative & Mass Citizen Scoring without consent in deviation of Fundamental Rights: Although opt-out can be an option for a citizen it might not always be realistic. For instance, what if I ask a bank to exclude me from their credit scoring but on the same time I maybe asking for a loan (e.g. mortgage) or apply for a credit card?	Requirements for Trustworthy AI, 8. Robustness: Do note that Robustness, Reliability and Resilience to attack (I call it Security) are non-functional quality characteristics according to the ISO 25010 Software Product Quality standard. Since AI systems are essentially software systems there can be a reference or relation of the current document to this ISO standard. I also think that it would be good if it can be defined which of the requirements concern the technical aspect of the algorithm and which ones the organisation that is using it. Also to me it seems that requirements such	KEY GUIDANCE FOR REALIZING TRUSTWORTHY AI: "Strive to facilitate the auditability of AI systems, particularly in critical contexts or situations. To the extent possible, design your system to enable tracing individual decisions to your various inputs; data, pre-trained models, etc. Moreover, define explanation methods of the AI system.". It might be interesting to pinpoint the value of externally developed explanation methods of an AI system as they can be more objective, reflect the state-of-the-art in the field and might help create a benchmark of how well an AI system can be explained.	The guidelines presented in this document are meaningful and clear. I also like that they're not exhaustive which gives some freedom, flexibility to use and extend them. Also, they do provide a guidance to AI producers and users for a fair and sustainable AI. As there are several initiatives towards governing AI (from IEEE, ECP's AIIA from or TRIM guidelines for financial institutions) and it would be interesting if this document hereto was explaining how it stands in relation to them and what makes it different.

						as Data Governance and Transparency can fall under the Accountability one.		Also, it is hard to imagine how these guidelines are going to be operationalised from corporations. For instance, a financial institution doesn't necessarily concern to be ethical but instead minimising the risk of bad creditors. That said, I expect most corporations to be interested in controlling their AIs but not necessarily invest in their ethics, unless they are either obliged by a regulation or they suffer from an incident and then they might start doing so.
Anonymous	Anonymous	Anonymous						This document is absurd in its use of terminology as it goes nowhere toward providing what is claimed. The term "trustworthy" mean non-interdependance, i.e. that the structure is secure irrespectively of actions. See e.g. terminology from PF7 security roadmapping http://blog.privacytrust.eu/public/Reports/securist-ab-recommendations-issue-v3-0_en.pdf Further the use of terms such as "ethics" and "trust" are 100% contradictory to the meaning of these terms. An "ethical" design would only be achieved when the design is trustworthy, i.e. inherently secure. Also the document works with assumptions of legal violations such as e.g. serverside face- or other biometrics recognition and secondary use of health data way beyond the possibility of informed consent and thereby ASSUMING non-secure system design. In short, the document is CLAIMING fundamental rights, trustworthiness, ethics etc. but going absolutely nowhere towards deserving or achieving such categorization. Especially the document ASSUME systemic violation of GDPR for the sake of profit and power structures that undermine European values. The document mention the terms such as "Privacy-by-Design" and "Security-by-Design" but ignore these and assume they are ignored in almost all other aspects.

Eberhard	Schnebel	European Business Ethics Network (EBEN) / Commerzbank AG	The Role and Function of the Ethical GuidelineA clear definition of what is a code aiming for helps to understand the function of ethics may provide and to avoid misunderstandings and misjudgements. This is important because the case of AI demonstrates in clear ways that ethical guidelines are mainly orientation and no legal framework. Leave the finding of law and the case law to the lawyers.1. Ethics has to inform willing people in how to proceed by developing and explicating technology into areas without clear notions. 2. Ethical Guidelines have to inform design processes for giving willing people the chance to design systems and codes in a way that they are subject to ethical ideas. And that they enable tracing individual decisions to the various inputs.3. Ethics may focus the fitting subject of AI development to inform citizens and community groups. If there is misinformation there is wrong judgement and misleading reservation and fears.4. Ethics has to keep Hayek's "development procedures" under competitive realm to engage in further development by limiting regulation on a minimal control grade.5. Ethics has to inform the "grey" area between a) strict rules, guidelines and law on one side, and b) personal judgement or social feelings and their social outcomes on the other side. So we need to focus this report more on this ethical side. 6. Ethics has additionally to inform lawmakers on how to set up fitting rules if social feelings are not able to serve social aims, like in dilemma situations (order ethics). 7. An ethical guideline itself is not able to prevent disfeature, wrongdoing and aberration. It may convey usually undocumentable intentions and impart values that are otherwise free of value, because they are very individual and subjective. Rights are rights and not values.Definition of Target AudienceThere should be a clear understanding of the target group of the guideline. It makes a big difference, if the guideline is for the work of public administration or for the work of private business.It is good to separate principles that are abstract high-level norms that developers, deployers, users and regulators should follow in order to uphold the purpose of human-centric and trustworthy AI. And it is helpful to refer on Values to provide guidance on how to uphold ethical principles. (p. 6) But this only works in the concretely specified applications and the fitting context of implementation. Private business and users may strengthen individual preferences but public use may avoid prejudices. Therefor we have to areas of application of AI-ethics:1. In Public Services an AI-Ethics-Guideline has to inform administration about democratic rules for implementation and application. 2. In Business an AI-Ethics-Guideline has to inspire AI-design as part of discovery procedures inside competitive settings. Implicated in AI design and use, different and often disparate stakeholder groups share some common values that can be used to strengthen further design coordination efforts.Rationale and Foresight of the Guidelines: Regulation?If there is no common understanding on the rationale of an ethical guideline there is no chance to create a fitting and operating framework.In societal contexts we will redefine and communicate again and again about what is human centric (is it democracy?). Therefore	I. Definition of the "common good" implies an idea of "summum bonum"There is no clear understanding of AI inside the dynamics of Social media. Here we need to address asymmetries first. In opposite to biomedical ethics, asymmetry is not a rational one than a behavioral one.What means "human centric" for an "ethical purpose" is not fixed. "Ethical" is related to communicative aspects of individual ideas of the application of the hidden ethical purpose. For assessment reasons this means we are not able to evaluate possible effects on this common good respectively. For utilitarian or consequentialist reasons we are not able to clearly define a common good on basis of a summum bonum. The development of AI itself is a service that enshrines two aspects in its mechanism and dynamics: Needs and desires! We are not able to fix a shared notion of "social feeling" (Hume) instead founding it based on the autonomy of ratio (with Kant, but this leads to several other misunderstandings). For AI-Guidance it is important to create a measure which is discovering itself, a measure that is developing itself inside designing and dealing with AI in social interaction, a continuously improving guidance.	II. The AI subject requires technical descriptions instead of "Respecting Fundamental Rights"Referring on fundamental human rights the Guideline loses all his accuracy. It does not explain why AI may destroy fundamental rights despite the fact, that users of AI may destroy them.Accountability: How are we able to create "statements of facts" or offence in new ways? Do we need to customize new circumstances on old offenses? An important distinction is between: a) Intrinsic societal concerns created inside AI development and b) What may AI obtain in criminal minds?Data Governance: Only good data sets bring us economic advantage. There is an intrinsic aim to use good data sets for economic reason. In this sense there are "requirements" that are very economic arguments for data governance, if only good data sets are able to bring an advantage for developers. To what goals is a proper governance of data and process related? What process and procedures were followed to ensure this proper data governance?"Design for all" and Fairness is as well an economic retention, if the design is made for a group of customers to satisfy their wishes, feelings and needs. Questions related are: What definition(s) of fairness is (are) applicable in the context of the system being developed and/or deployed?Privacy: Privacy is actually not an object of AI, because it does not actually require private data, but rather statistical data about behaviour patterns. Is the personal data information flow in the system under control and compliant with existing privacy protection laws?A major problem of AI ethics is the uncertainty of how data is transformed into business models. Some people think that it is the data itself and therefore fight for ownership of data and data sovereignty of user. However, it is rather the behavioral models that can be derived from data that lead to business models. This second version of "data oil" generates business value from data as it is processed rather than stored.Transparency: There is no conventional information asymmetry, but one that is in different use or term of information: AI may have some information on our behavioral pattern if trained. And additionally, there is a "Transparency Purpose": Public Service and Competition. Traceability: What measures are put in place to inform on the product’s accuracy? On the reasons/criteria behind outcomes of the product?Robustness: Trustworthy AI requires that algorithms are secure, reliable as well as robust enough to deal with errors or inconsistencies during the design, development, execution, deployment and use phase of the AI system, and to adequately cope with erroneous outcomes.	III. Requirements of explainable AI for assessing trustworthy AI If we take on an engineering approach there is no need for such things like "trustworthy" and "explainability". Here the Ethical Guidelines switch the "ethical" approach in a narrow sense towards a "principal" one. To announce clear and concrete requirements means an early constriction of dealing with a much wider ethical Issue. The guideline builds an engineering approach to set techniques instead of communication, to set induction and deduction instead of abduction.Because an "EU Ethics Guideline" is nothing that can be separated from administration and right, it creates covetousness to establish an auditable standard. We need to reduce the topics of assessment to avoid the limitation of guidelines on this narrow sense. Following I will concentrate on meaningful notions of Assessment topics:Accountability: Are the skills and knowledge known and present in order to take on responsibility in AI design and training eg by an ethical oath that highlights virtues and duties?Ethical Review Board: Has an Ethical AI review board been established that is aware of how it can give people an inner orientation? Are there mechanisms of communication to discuss grey areas?Governing AI autonomy and Human control: Is a process foreseen to allow human control, if needed, in each stage? Within the organisation who is responsible for verifying that AI systems can and will be used in a manner in which they are properly governed and under the ultimate responsibility of human beings?Non-discrimination: What are the sources of decision variability that occur in same execution conditions? Does such variability affect fundamental rights or ethical principles? How is it measured? Is a strategy in place to avoid creating or reinforcing bias in data and in algorithms? Respect for (& Enhancement of) Human Autonomy: Do users have the facility to interrogate algorithmic decisions in order to fully understand their purpose, provenance, the data relied on, etc.?Resilience: Resilience is a technical issue; Reliability is a fact of competition and of complexity; Sensibility and accuracy is subject to the use of AI in terms of product quality. What other data sources / models can be used to eliminate bias? Fall back plans are nonsense.Safety: Have the potential safety risks of (other) foreseeable uses of the technology, including accidental or malicious misuse thereof, been identified?	General remarks1. There is no clear specification of function of the ethical guideline. It should not be an administrative set of guidelines, but a motivation to create AI and communicate on ethical values related to AI. The pillar three of the commission “ensuring an appropriate ethical and legal framework to strengthen European values” moves the objective of an ethical guideline and a legal framework very closely together, what is not appropriate in the case. We should separate ethical framework from legal framework. Whilst the latter concentrate to design matter of facts, the former focus on guidance for orientation without boarder. Leave the finding of law and the case law to the lawyers.2. The missing specific definition of target audience (either public administration or private business) is a big source of dilution in this case. For public institutions and administration it is important to be transparent. For private business we always fight for hidden reasons as part of privacy and competitive advantage. We therefor established other rules, like consumer protection, product liability or antitrust law, what has to be specified and further developed to cope with AI solutions.3. No clear definition of the object of the guideline: Is it specific for AI - then there are conventional concepts such as technical failure, user liability, etc. completely sufficient – or is it an unknown thing with digitalization in the farer future – then the guideline is not relevant at all. The definition of the object sets the focus on technical descriptions.4. The rational, implications and foresight of the guidelines, which are initially of a voluntary nature, cannot currently be assessed. According to a cursory review, the high requirements for transparency, explainability, comprehensibility and non-discrimination of AI systems in particular could have a negative impact on broad economic use. What are topics of regulation and what are topics of ethics-communication?5. There is no adequate adaption of data-issues and Deep Learning for shaping explainability and trustworthiness. It is difficult to identify and qualify when a model is homogeneous or when it is biased. There is still some work to be done on data quality - a clear metric that shows how the data set affects the model.6. No clear adjustment to concrete ethics approach. Is it a modern ethics of communication and virtue or an obsolete natural rights issue? Is it utilitarian? Do they focus on a common good and a “summum bonum” of the society?7. Ethics Guidelines are not a solution but the beginning of a process and discussion and should be separated from all legal or soft-law ambitions.Questions about the general administrative methods of AI-GuidelinesIs an Ethics Guideline really something that can be separated from administration and right? Or is it an administrative tool? >> Then we should distinguish between administrative Ethics and societal Ethics and their respective guidelines. The EU therefor has to clearly line out their administrative approach.• Is the EU able to set something different from administration that channels and motivates?• Is AI something that can be administered on the actual stage of development? Whose design can be structured? Do we need a very new kind of administration for fast developing AI? We

				it is not possible to adopt biomedical principles towards AI, where we use a very assistive or health support approach and evaluate interventions in biological settings. What is boosted by AI? (8) For AI it is important to focus on the goal of establishing ethical dynamics on shaping societal communication and social media, the grey area.Officials and administration always argue creating some soft law by defining ethical principles. By doing this, they lose some measure of the grey area between "rights" and "arbitrariness". If we accept the dilemma of a summum bonum, ethical values and principles may have a very dynamic and temporary character to frame human relationship and dynamics of communication.Identification without consent: Personal identification leads to focus on irrelevant AI mechanisms, because persons are not goals for calculation but are features or behavioral pattern.Covert AI-Systems and Autonomy: We don't have to be afraid of covert AI systems if we just make sure that our autonomy is maintained. This will happen sooner by refraining from monitoring by AI and focusing on using assistive AI systems. (see: Zuboff Shoshana who discusses much more subtle mechanisms of undermining autonomy by social media.)Questions:Is an "Ethical Guideline for AI" really good in taking the ideas of established techniques?Is AI still a very emerging technology, permanently designed and developed, with unknown use cases and outcomes, permanently involved in procedures of discovery?What will change is the AI development process is part of the competition as discovery procedure?		need two separate ways of administration:1. Administration of clear and stable issues2. Administration of development and discovery inside societal dynamics• Is Design something that should again be structured differently, not with instruments of contemporary administration of established technologies?• Is administration something that collects knowledge on structuring societal frameworks?Not carved out:Competition: What is the value and are the opportunities of competition as motivation for discovery and innovations?Complexity: Reproducibility, reliability and comprehensibility are not possible.Behavioral Biases and Marketing:Governance: There is a major concern to govern issues without clear goalsDigital Development and AI-Development as Discovery Procedure: No Goals, No Summum Bonum for the adjustment of a clear target image
Andrés	Abad Rodríguez		In point 4, I would add a new principle: Principle of privacy: "Do not disclaim more than what you were allowed to(if any)" AI will use an immense amount of data, including EU citizens private information. Despite making that data anonymous, people must be able to allow and remove permissions at any moment on their data. This is especially important with sensitive data such as medical records.		I would add a point about "validation": - When is a system good enough for being used in production? - Who can asset the systems? An AI EU regulator? - How could be some minimal KPIs be identified for considering a system good?	Thanks a lot to the HLG for the document. It is a very good starting point. Please, do not hesitate to contact me for more information if needed.
Leon	Kester	TNO		My comment consist of two main points: 1 Utility based ethical goal functions are also essential beyond level 3 In the guidelines a five level model is used for intelligent systems where the use of utility functions is only considered useful up to level 3. Further, in the paragraph on trustworthiness the phrase “rules, norms and laws that govern AI” suggests that utility functions are not useful on the higher levels. As researchers working on the design of intelligent autonomous systems, however, we experienced that a normative approach for governing AI (describing rules, norms and laws for every possible state-action combination) is not feasible for realistic scenarios due to ‘state-action space explosion’ (Werkhoven et al., 2018). Actually this is something that was already shown by Asimov with his robot laws. Utility based reasoning is also essential at levels 4 and 5, to achieve safe and secure systems which behave in an ethical way acceptable by society (Aliman and Kester,		The use of the term ‘trustworthy’ can be interpreted in two ways: - The intelligent system is inherently safe, secure and optimizes on ethical goal function as provided by humans, therefore it deserves our trust for mathematically comprehensible reasons. - Due to the behavior of an intelligent system e.g. by working with such a system the human tends to trust it. The first interpretation reflects what we should want as a society while the latter is in my opinion an anthropomorphic view which could become highly dangerous with a further advancement of AI development. By way of example, one could e.g. consider the movie ‘Ex Machina’ where an intelligent system acted in a way that it was perceived as being ‘trustworthy’ according to the second definition (and not according to the first one). This emotionally biased erroneous assessment had fatal consequences.

							2018). In order to do so, a utility function - which we call an ethical goal function* - needs to be specified that acts on levels 4 and 5. We are well aware of the objections to consequential ethics/utilitarianism but are confident they can be resolved. Moreover, in our opinion they have to be resolved because we see no technically feasible alternative. 2 Responsibilities of manufacturers and legislators must be separated Because of the orthogonality of the problem solving capability of AI and its final goals as formulated in an ethical goal function, the responsibility of the manufacturer/programmer and the legislator can and should be clearly separated. It is strange to expect from manufacturers/programmers to be ethical while the legislator is not able to formulate what that actually is. Manufacturers themselves look for clarity from governments. For instance, Google even stated that** "Some contentious uses of AI could have such a transformational effect on society that relying on companies alone to set standards is inappropriate — not because companies can't be trusted to be impartial and responsible, but because to delegate such decisions to companies would be undemocratic". So in general government/legislators should be responsible for specifying ethical goal functions while the manufacturers/companies should be responsible for making safe, secure systems that optimize on ethical goal functions. Thereby, an urgent action of governments is needed on an international level with regard to these issues. *https://www.tno.nl/en/tno-insights/articles/does-ethical-artificial-intelligence-exist/ **https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf References: Werkhoven, Peter, Leon Kester, and Mark Neerincx. "Telling autonomous systems what to do." Proceedings of the 36th European Conference on Cognitive Ergonomics. ACM, 2018. Aliman, Nadisha-Marie, and Leon Kester. "Hybrid Strategies Towards Safe "Self-Aware" Superintelligent Systems." International Conference on Artificial General Intelligence. Springer, Cham, 2018.		
Paolo	Giudici	Professor of Statistics, University of Pavia					Page 26 - assessing robustness - accuracy through data usage and control:I suggest to insert a third bullet point to make accuracy evaluation more effective and measurable, as follows:"For each of the considered accuracy measures, provide a threshold, either deterministic (cut-off point) or stochastic (critical value or p-value) above which accuracy is attained"Pag 27- assessing traceability - methods of testing the algorithmic system, in case of a learning based model, I suggest to insert a test of the statistical data analysis model employed, and not just on the data used, as follows:"information about how the data is analysed should be provided, including: pre-processing of the data (data cleansing, outlier detection, missing data treatment,	I like very much the structure of the document, and the rationale behind it

							organisation of the data); analysis of the data (descriptive and predictive statistical models employed, which algorithm/software has been used to implement them); model validation (model comparison tools, model selection criteria). This for both statistical learning models (more transparent, less efficient) and machine learning models (more efficient, less transparent)	
Sjoert	Fleurke	Radiocommunications Agency Netherlands (Agentschap Telecom)	-	General: Be careful in demanding that AI has almost an ethical purpose in itself. Otherwise one will run into the difficulty that any 'lower' design components, and any algorithm within AI must prove itself to support only the positive, i.e. doing good. This will be impossible and it will be contra-innovative. It should be considered to treat AI as a set of complex means, tools, methods, software, algorithms etc. where measures should be taken to ensure that users, manufacturers, internet companies, etc. use them properly. This secures the right distinction between actor and tool, subject and object. It is about keeping humans (and/or legal companies) responsible for what they do with the AI-tool. It is keeping humans always responsible and accountable over the technology they invent, from the creative development stage up to the usage stage. In case AI gets rights above humans, there may be tremendous difficulties in our juridical system (i.e. subject/object). ad 5.3, page 12: We agree that opt out options are important. But instead of clicking all the time "I accept", like todays "I accept cookies", we must help to create for each citizen his/her own digital profile (a digital twin) in which only this person can enter his/her own principles and preferences, so that this profile helps ignoring AI–using options, AI-using methods, or AI-using companies a person does not like to work with. Ad 5.5 page 12 : Our suggestion is to conduct impact assement(s) at certain stage(s) to consider the need of new laws or new regulatory measurements (like e.g. monitoring or auditing system and/or law enforcement)	-	In addition to the 10 requirements listed in this chapter we have 2 possible additions: 1. We would like to argue that the data used to train the AI system must have been gathered in a legal and ethical way. If this is not guaranteed, it will result in rejection of the AI system by (a part of) the public. 2. Furthermore, because developments in the field of AI go rapidly, we would like to have added that AI system operators have an obligation to maintain and update their algorithms/methods regularly in order to guarantee that the outcomes are more or less as reliable as possible at that moment in time. In case no such obligation exists, this may lead to the use of obsolete (and cheaper) methods which may reduce trust in AI by the public.	We underline the importance of transparency in the introduction of AI as a whole, and specifically in the way algorithms and machine learning include/exclude people or people's preferences.
Anonymous	Anonymous	Anonymous						

MARIO	ROMAO	INTEL	We fully agree with the basic tenet of the guidelines as expressed in this section that "AI holds the promise to increase human wellbeing and the common good but to do this it needs to be human-centric and respectful of fundamental rights" and "that no legal vacuum currently exists, as Europe already has regulation in place that applies to AI."We also agree that protecting individuals and their data goes beyond legal compliance requirements: it means embracing societal values and working to build a much needed trust in the technologies and their impact to people.	2. From Fundamental Rights to Principles and ValuesInformed consent has been a valuable tool to empower citizens and give them control over data but has always had limited effect due to the tremendous burden it places on individuals to understand how data is collected, processed, used. The legitimate interest of the entity processing data should be balanced against the legitimate expectations of the individuals, so that it can supplement consent where context is appropriate. This would work in concert with substantial protections to individuals and obligations on organisations (e.g. accountability approaches).In this section, as the concept of informed consent traditionally belongs to the data protection sphere, we suggest to clarify how this would apply to the ethical dimension which includes privacy but is broader.3.1. Respect for human dignityIn describing the mapping of respect for human dignity onto AI, we suggest utilising a language that ensures that "respect", without seeming to exclude more mundane applications and that at the same time upholds those principles (e.g. AI in entertainment):"To specify the development or application of AI in line with human dignity, one can further articulate that AI systems are developed in a manner which upholds humans' physical and moral integrity, personal and cultural sense of identity as well as the satisfaction of their essential needs."3.4 Equality, non-discrimination and solidarity including the rights of persons belonging to minoritiesWe suggest to rephrase the sentence "Equality also requires adequate respect of inclusion of minorities, traditionally excluded, especially workers and consumers" as its current meaning does not seem clear (excluded from the workforce and market economy?).3.5 Citizens rightsThe suggested right of a citizen to "systematically be offered to express opt out" of "automated treatment of their data by government bodies" seems not to be compatible with current practices (e.g. in healthcare, taxation). The emphasis should be rather on ensuring adequate technical safeguards in the automated decision making process by government bodies such as de-identification techniques and strong encryption (including homomorphic encryption) as well as a sound legal basis to institutionalise those automated practices in specific and well-identified contexts. Where possible, substantive individuals' rights (to opt out, to appeal and to be informed) should be also made available.4. Ethical Principles in the Context of AI and Correlating ValuesThe Principle of Non maleficence: "Do no Harm"We suggest the following change: "Harms can be physical, psychological, financial or social. Amongst others, AI specific harms may stem from the treatment of data on individuals (i.e. how it is collected, stored, used, etc.). In this specific case, to avoid harm, data collected and used for training of AI algorithms must be done in a way that avoids discrimination, manipulation, or negative profiling." As there may be harms deriving from factors other than data.We also suggest a new text for footnote 12 under environmental awareness. As the reference only highlights social challenges, we have modified it to also reflect environmental challenges. There has been a lot of media attention on sourcing of cobalt and lithium from regions such as	1. AccountabilityThe current description relates more to liability than to accountability. We suggest the following wording:"Effective AI governance should include accountability measures, which could be very diverse in choice depending on the goals. Accountability can be described as the ability to demonstrate that appropriate measures have been put in place by an organization to minimize risks identified for the specific AI system and usage. These technical or organizational measures should be tailored based on each business' needs as well as the specific risks themselves. Consequently, regulators could deem accountability measures as a mitigating factor in case of incidents. Mechanisms for compensation can range from monetary compensation (no-fault insurance) to fault finding, to reconciliation without monetary compensations. The choice of compensation mechanisms may also depend on the nature and weight of the activity, as well as the level of autonomy at play. An instance in which a system misreads a medicine claim and wrongly decides not to reimburse may be compensated for with money. In a case of discrimination, however, an explanation and apology might be at least as important."6. Respect for (& Enhancement of) Human Autonomy We suggest rephrasing "AI products and services, possibly through 'extreme' personalisation approaches, may steer individual choice by potentially manipulative 'nudging'. At the same time, people are increasingly willing and expected to delegate decisions and actions to machines (e.g. recommender systems, search engines, navigation systems, virtual coaches and personal assistants)." as mere examples:"Examples of AI systems that might have an impact on human autonomy are 'extreme' personalisation approaches, which may steer individual choice by potentially manipulative 'nudging', and the fact that people are increasingly willing and expected to delegate decisions and actions to machines (e.g. recommender systems, search engines, navigation systems, virtual coaches and personal assistants)."10. Transparency We agree with the statement, earlier in the section "Scope of the Guidelines" that "While the Guidelines' scope covers AI applications in general, it should be borne in mind that different situations raise different challenges. AI systems recommending songs to citizens do not raise the same sensitivities as AI systems recommending a critical medical treatment". This statement would also apply to transparency requirements. Not all uses of AI require the same level of scrutiny or "transparency". Therefore we suggest the following changes:"Transparency concerns the reduction of information asymmetry. Explainability – as a form of transparency – entails the capability to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environments, as well as the provenance and dynamics of the data that is used and created by the system. Being explicit and open about choices and decisions concerning data sources, development processes, and stakeholders should be required for those AI systems involving high-stake decisions (those having legal effects or otherwise significantly affecting human beings). In particular, governments should determine which AI	Respect for Privacy – The question "is the system GDPR compliant" is probably too EU focused, while "Is the personal data information flow in the system under control and compliant with existing privacy protection laws?" would promote the international convergence on similar and compatible privacy approaches. The question about informed consent would more appropriately be formulated as: - How can users seek information about the legal basis for processing personal data?And we would find it useful to include:- Have accountability approaches been adopted, such as technical mitigations (e.g. data de-identification techniques, differential privacy, encryption) or organisational measures (e.g. training, internal audit)?

Africa and South America. These two minerals are important for battery production and supply chain abuses have been uncovered. Procurement is a tool that can be used to address these challenges. The suggested text is as follows:"Items to consider here are the positive and negative environmental impacts of computing power to run AI systems and the application of voluntary Data Centre initiatives such as the EU Code of Conduct to optimise operation within these facilities, and the procurement of minerals to fuel the batteries needed for all devices involved in an AI system. For the latter, it is important that minerals such as cobalt and lithium are sourced responsibly from conflict-affected and high risk areas in order to avoid abuses in supply chains such as forced labour and mercury pollution at mines."The Principle of Autonomy: "Preserve Human Agency"The text seems to postulate the right to opt out and to withdrawal from every single AI use case. Instead, we suggest these rights should be set by the legislator/regulator whenever deemed necessary and in specific use cases. The Principle of Justice: "Be Fair"Effective redress, as introduced in this section, should not be presented as a quality derived from the fact that an AI system is or not in place. It is rather a quality derived from a legal right or judicial procedure applied to a wrong, independently of the technology used. In other words, AI systems should not add to or subtract from the redress rights stemming from the implementation of the law and judiciary proceedings. At the same time, defining multiple avenues to guarantee to all citizens access to judicial redress would be of paramount importance. The Principle of Explicability: "Operate transparently"We agree that "Transparency is key to building and maintaining citizens' trust in the developers of AI systems and AI systems themselves". However, the use of the term "business model transparency" in this section is not the most adequate, as that is broadly used in contexts linked to financial and non -financial reporting obligations, the level of corporate disclosure needed to attract external investments and in general, the level of information considered adequate for customers, investors, suppliers and employees to engage in a company's activities.In the present case, where the focus is explicability and transparency of AI systems, we suggest removing the reference to "business model transparency" and rephrasing the first paragraph of this section as: "Transparency is key to building and maintaining citizens' trust in the developers of AI systems and AI systems themselves. Transparency implies that human beings are knowingly informed when AI systems are in use, their purpose and that those are auditable, comprehensible and intelligible at varying levels of comprehension and expertise, and according to the level of risk involved." 5. Critical concerns raised by AIWhilst these may be valid concerns, at this stage, and as the Executive Summary states, the goal of the guidelines are to offer guidance on the concrete implementation and operationalisation thereof into AI systems. In Europe, the focus region of these guidelines, the issues presented (Identification without consent; Covert AI systems; Citizen Scoring; LAWS) are either plainly unlawful or would have to be in accordance with EU or Member State implementations need algorithmic explainability to mitigate discrimination and harm to individuals.

								legislation to be allowed. Also, mitigation strategies have been addressed elsewhere in the guidance (e.g. transparency as a solution to address covert AI systems). The issues presented in section “5. Potential longer term concerns” are purely speculative. We thus recommend to remove section 5. altogether from the guidelines, continue the debate on these and eventually include them as an annex for “further work”. These would meet the statement in the Executive Summary that “Moreover, the Guidelines should be seen as a living document that needs to be regularly updated over time to ensure continuous relevance as the technology and our knowledge thereof, evolves.” while keeping its core objective of being concrete and operational today.
Wayne	Grixti	Malta Digital Innovation Authority			Re Paragraph 5.5 - Potential longer-term concerns: MT agrees that whilst these may be speculative and belonging to the distant future, one needs to consider that development of AI systems (and consequently the concerns which this gives rise to), may go beyond what we currently are able to anticipate, or what currently appears realistic, and indeed may come upon us much faster than we think.		MT feels that a due diligence exercise in relation to the developers of an AI system, and possible the investors and owners of the AI system, may be useful. In certain specific cases, this may also be considered for the users and purchasers of a particular AI system. Furthermore, in relation to any information that is required to be provided to users, it may be useful to assess whether such information is easily accessible and intelligible.	It is good to see that the Guidelines do not stop at a list of values and principles for AI, but rather offer practical guidance for “implementation and operationalization” of such principles into AI systems. The Guidelines make a good point regarding the fact that generating trust in AI (through an ethical approach) will facilitate broader uptake of AI and promote responsible competitiveness.i. On Page ii of the Executive Summary it is stated that “The Guidelines are not meant to stifle AI innovation in Europe, but instead aim to use ethics as inspiration to develop a unique brand of AI, one that aims at protecting and benefiting both individuals and the common good.” The Guidelines further make the point that what benefits individuals sometimes goes against the common good, and this creates tension between the two differing positions that an ethical framework is intending to protect. MT agrees that this is inevitably going to be the case on a regular basis, and the fact that the common good will often prevail over an individual’s rights is a fact that needs to be accepted.ii. On Page ii of the Executive Summary it is stated that “Trustworthy AI has two components: (1) it should respect fundamental rights, applicable regulation and core principles and values, ensuring an “ethical purpose” and (2) it should be technically robust and reliable since, even with good intentions, a lack of technological mastery can cause unintentional harm.” On Page ii of the

							Executive Summary it is further explained that "ethical purpose" means "reflective of, fundamental rights, societal values and the ethical principles of Beneficence (do good), Non-Maleficence (do no harm), Autonomy of humans, Justice, and Explicability." MT agrees with these principles as being the fundamental building blocks of AI in Europe.iii. The Executive Summary states that "Finally, beyond Europe, these Guidelines also aim to foster reflection and discussion on an ethical framework for AI at global level." We feel that the EU should perhaps do a bit more in this respect, by actively instigating discussion of these matters on a global level. Policies at EU level will have limited effect in protecting EU citizens if other countries such as China and the US, who are far more advanced than Europe in terms of AI development, are not restricted by at least the same high-level principles.iv. MT also agrees that special attention needs to be given to "situations involving more vulnerable groups such as children, persons with disabilities or minorities, or to situations with asymmetries of power or information, such as between employers and employees, or businesses and consumers."v. MT also stands with, and look to support, the EU Parliament's resolution of 12 September 2018 and all related efforts on Lethal Autonomous Weapon Systems (LAWS). vi. MT feels that more in-depth discussion of the potential of AI uses for terrorism could be included, as this is a very real concern: by way of example, one could consider, for example, the hacking of autonomous vehicles and the disaster such a situation is able to bring about. MT feels that further studies should be carried out in this respect (although this could be intended as part of AI HLEG's second deliverable, due in May.
Maria Luisa	Stasi	ARTICLE 19	• ARTICLE 19 contests the major role that the draft attributes to ethic and urges to replace it with an approach focusing on the protection of human rights. In the section dedicated to "The Role of AI Ethics", the draft states: "The goal of AI ethics is to identify how AI can advance or raise concerns to the good life of individuals, whether this be in terms of quality of life, mental autonomy or freedom to live in a democratic society." ARTICLE 19 believes ethics is insufficient to protect individuals from the harm that AI can inflict on them. Therefore, the role of ethics is minor, and certainly less important than the role that legal provisions shall have for this purpose. • The scope of the guidelines does not include an awareness of dual-use AI technologies. However, in many instances this kind of technologies raises tremendous challenges for individuals' fundamental rights. Various countries have already in place norms regulating the import/export of these technologies, and some companies are already calling for States to provide regulatory guidance concerning development. ARTICLE 19 believes that the guidelines shall duly consider dual-use AI explicitly and establish rules aimed at creating safeguards and guarantees for individuals' rights.	• ARTICLE 19 that the session on "Fundamental Rights and Human Beings" could be more efficient if an explicit mention on specific Articles of the Treaties and of the Charter would be included where relevant. This will enhance legal certainty and facilitate the identification of relevant case law where needed. In particular, section 3.2 could include explicit reference to the rights and freedoms that compose the "freedom of individual", such as, for example, freedom of expression, privacy, non-discrimination. • Section 3.4 raises challenges. Various AI systems are behind intellectual property protections, how can one guarantee access to information and knowledge? ARTICLE 19 believes that sometime distinctions can be legitimate. In addition ARTICLE 19 notes that workers and consumers are not "minorities". • ARTICLE 19 suggests that the citizens' rights mentioned in section 3.5 shall be equally guaranteed where data are used for automated decision making by both States and private actors. • Session 4 on "Ethical Principles in the Context of AI and Correlating Values" contains a list of high level principles: beneficence, non-maleficence, autonomy, justice, explicability. The majority of these principles have been already established and made applicable by different branches of EU law, such as the General Data Protection	• ARTICLE 19 suggests that section 1 should focus on how to establish accountability, and on who is to be held accountable, rather than on possible remedies. Possible remedies could be suggested in a separate session to be added. • With regard to section 2, dealing with data governance, ARTICLE 19 warns on the fact that the efforts to combat bias shall in no case lead to the violation of fundamental rights. In addition, on issues concerning collection and purpose limitation, ARTICLE 19 recommends to insert references to the relevant provisions of the General Data Protection Regulation. • Concerning section 3, ARTICLE 19 notes that not all systems are, or should be, intended for all. ARTICLE 19 therefore suggests having a qualifier, which makes explicit that the State-deployed systems intending to serve all should allow all citizens to use the products or services, regardless of their age, disability status or social status, or better, in a way that is not discriminatory to protected attributes. • The section dedicated to good governance shall include remedies, redress mechanisms and due process guarantees. ARTICLE 19 also considers that guidance about which kind of remedies companies could design would be useful in providing legal certainty.	No comments on this part.	• The guidelines focus on ethics, rather than on fundamental rights and existing legal frameworks. The draft suggests a new concept of "ethical purpose" that should include fundamental rights, principles and values. Nevertheless, ethics and law remain two different concepts. ARTICLE 19 believes there is no added value in introducing a nebulous concept like "ethical purpose" which, on the contrary, appears to create confusion and undermine existing legal rights and duties. • The draft explicitly takes into account privacy, non-discrimination and human autonomy, and dedicate to them specific provisions. Nevertheless, there are other fundamental fights, such as freedom of expression that might be strongly affected by AI. ARTICLE 19 calls for an explicit mention that all fundamental rights affected by AI deserve adequate protection. • The draft guidelines are based on the concept of "Trustworthy AI". ARTICLE 19 notes that trust is a relationship between peers, where the trusting party, while not knowing for certain what the trusted party will do, believes any promises being made. Therefore, one cannot trust AI, and AI cannot be trustworthy. For AI, accountability has to be used. AI can be accountable. Trustworthiness should be for institutions that hold AI accountable.

Regulation. ARTICLE 19 believes that a vague reference to them risks undermining their legal status, diminishing legal certainty and suggesting somehow that the enforcement of these principles depends on the inclinations or goodwill of companies that develop AI. Instead, reference should be made to the relevant rules that guarantee these principles and/or implement them.
• Session 5.1 "Identification without Consent" states: "Differentiating between the identification of an individual vs. the tracing and tracking of an individual, and between targeted surveillance and mass surveillance, will be crucial for the achievement of Trustworthy AI.". ARTICLE 19 calls for a deeper analysis of the distinctions, as well as of the safeguards and the possible shortcomings. More in general, ARTICLE 19 notes that AI identification technologies interfere with privacy, right to anonymity and potentially freedom of expression of individuals, and recalls that the interference shall be subject to the three-party test of legality, proportionality and necessity.
• ARTICLE 19 calls for the inclusion of the right to know when one is subject to a decision/interaction with AI in session 5.2.

• The section dedicated to "Traceability and Auditability" (p.22), both elements are presented as a mere option. ARTICLE 19 believes this call shall be reinforced. In addition, ARTICLE 19 suggests to add interpretability among the parameters.

• In the section dedicated to "Regulation", ARTICLE 19 calls for an explicit reference to due process when dealing with redress and remedies.

• ARTICLE 19 believes that the section dealing with "Standardisation", shall contain the recommendation that standard setting bodies include human rights impact assessment in their considerations.

• The text repeatedly refers to "data subjects". ARTICLE 19 urges the HLEG to remember that data subjects are human beings with all their rights as guaranteed by EU Charter. Therefore, when discussing about safeguards for data subject, the approach has always to be: do the planned legal provision/remedy adequately guarantee the protection of people's fundamental rights?

• In the executive summary, it is states that "...on the whole, AI's benefits outweigh its risks, (...).". This assumption is not proven, and it should not be taken as a reason to relax the approach on regulating AI.

• The draft contains numerous expressions which indicate that compliance with what prescribed or recommended is left to the goodwill of AI developers. This interpretation is corroborated by the focus on ethics rather than on legal obligations. Examples are, among many: "Moreover, keep those requirements in mind when building the team to work on the system, the system itself, the testing environment and the potential applications of the system" (p.3). "Keep those requirements in mind" is a weak expression that does not create any duty on the developers. "Strive to facilitate the auditability of AI systems, particularly in critical contexts or situations." (p.3). "Strive to" guarantees too large margin of manouvre for developers. "Be mindful that there might be fundamental tensions between different objectives" (p.3). Being mindful does not imply the requirement for any specific action. The draft suggests to "Communicate and document these trade-offs": however, it is not clear to whom and why they should be communicated, nor how they should be documented. All in all, the softness of this language undermines legal certainty. ARTICLE 19 urges to opt for a stronger approach and to avoid vagueness around what companies are called to do, not to do, or to comply with.

• The draft states "This guidance forms part of a vision embracing a human-centric approach to Artificial Intelligence, which will enable Europe to become a globally leading innovator in ethical, secure and cutting-edge AI." (p.3). ARTICLE 19 strongly believes that Europe shall commit to lead on a human rights friendly AI, not on an ethical one. On this regard, ARTICLE 19 recalls that the EU Charter of fundamental rights establishes the obligation, for the EU institutions as well as for the member States when applying EU law, to guarantee full respect of the fundamental rights listed therein. This remains valid concerning any actions that EU institutions undertake with regard to AI.

• The concept of "human-centric" approach is based on human values (p. 4). ARTICLE 19 urges this concept to be based on human rights instead, and to therefore take into due account the existing international legal framework to protect them.

Niklas	Horstmann	Deutsche Telekom AG	- Our understanding of the guidelines: European values, enshrined in digital ethics, are a competitive advantage for the development of AI and enablers of progress and trust. Our understanding is that the guidelines are intended to contribute to leveraging this potential. We understand that the guidelines neither constitute nor directly prepare new regulation regarding AI. A tightened legal framework would be detrimental to the European AI ecosystem and thereby constitute a societal disadvantage. In this case, the normative forces of technological developments from West and East would ultimately determine the factual rules. Europe needs a strong AI infrastructure to be perceived as a global player and to exploit all value creation potentials of this new technology.- Rationale (p. 1): To foster support for the guidelines and facilitate their impact in common practice of AI development, deployment and use, the need/reasons for AI guidelines (e.g. establishing a European approach to AI in order to ensure European competitiveness) should be better motivated (cf. p. iv).- The Role of AI Ethics (p. 2): We strongly support the idea that the guidelines' scope is digital ethics with a focus on AI technology and the questions arising with regard to its impact the entailing responsibility of all stakeholders. The ambition should be to bring the guidelines fully into practice as/via domain-specific ethics code(s).- Endorsement mechanism (p. 2): The introduction of a mechanism under which stakeholders will be able to "formally endorse" (p. 2) the guidelines raises questions regarding its practicality: What are the consequences of an endorsement? Would this (fully or partly) replace self-initiatives such as codes of conduct or self-binding guidelines? Would signatories thereby fall under specific external governance/auditing? And wouldn't non-signing of these guidelines artificially create the impression that a stakeholder does not support ethical considerations regarding AI? Lastly, it appears difficult to achieve broad endorsement of the guidelines in the form of a 'take-it-or-leave-it' approach, where some guidance might be considered acceptable by signatories, and others might not. While the intention to regularly update and evolve the Guidance by treating it as a "living document" (p. iv) is comprehensible, it might also lower stakeholders' willingness to endorse it formally.- Trustworthy AI (p. 2): We agree with the assessment that "no legal vacuum currently exists, as Europe already has regulation in place that applies to AI" (p. 2), not least due to the technology's cross-sectoral nature. While the guidelines are not intended "as a substitute to any form of policy-making or regulation" (p. 3), the before-mentioned conclusion nevertheless must be taken into account for the HLEG's second deliverable, i.e. the AI Policy and Investment Recommendations, due in May 2019. In this context, we suggest a footnote clarifying that due to fast technological developments the existing legal framework may need to be further developed and adapted to potential new requirements, such as with regard to cyber security, information security or competition law. In particular, competition law should be equipped with the necessary tools to intervene in cases of market abuse related to exclusive access to data and platforms and to address emerging	- The EU's Rights' Based Approach to AI Ethics (p. 5): The guidelines should specifically carve out the EU understanding of terms like "wellbeing and the common good" (p. 5) in order to ensure clear differentiation from alternative interpretations of these terms around the globe. In particular, it must be clarified that wellbeing, common good and beneficial AI are to be seen under the lens of civil rights. It is not sufficient to state that we want to follow the "chapters in the Charter" (p. 5).- Informed consent: At numerous places in Chapter I (e.g. sections 2, 4 and 5.1), the notion of "informed consent" is being addressed and interpreted as the value to operationalize the principle of autonomy in practice. Further, informed consent is put forward as the needed requirement to ensure "explicability and non-maleficence" (p. 10). While consent can be one solution to guarantee accountability and transparency towards users, it is by far not the only one. E.g. in reference to the GDPR, processing of personal data for the purpose of offering an AI-based service can be based on 6 different legal bases (consent just being one of them), such as processing necessary for the performance of a contract or for legitimate interest. In addition, the principle of compatible further processing (Article 6(4) GDPR) allows companies to use personal data for other than the initial purposes without the need for an additional legal basis. Consent is thus not the sole value and solution to enhance explainability. E.g. voluntary approaches such as a "one-pager" that explain in simple terms for what purpose personal data is being collected, can enhance transparency much better. Therefore, the notion of informed consent is given too much prominence in this guidance and misleads it to be only and best requirement to preserve autonomy and explainability.- Equality, non-discrimination and solidarity including the rights of persons belonging to minorities (p. 7): AI development, deployment and use should adhere to the fundamental right of equal treatment, as set out e.g. in Chapter III of the EU Charter of Fundamental Rights. In this context, the document states that "equality of human beings goes beyond non-discrimination" (p. 7). To avoid misconceptions, conflicts with the fundamental right of individual freedom, or the notion of a 'levelling down', this statement should be revised to capture more precisely what it is supposed to mean in the context of AI.- Vulnerable demographics (p. 9): While Chapter I explicitly states that particular attention should be payed to vulnerable groups (e.g. children, asymmetries of power of information), no detailed explanation, input or background is given to this guidance and how it should be observed in practice. It is however essential to explore in more detail the implications to such guidance, e.g. regarding business-consumer relationships or situations where AI-driven business models are not solely offered to one specific vulnerable group (but users in general).- Principle of Autonomy (p. 9f): As stated in footnote 13 ("This includes a right to individually and collectively decide on how AI systems operate in a working environment. This may also include provisions designed to ensure that anyone using AI as part of his/her employment enjoys protection for maintaining their own	- Requirements of Trustworthy AI (p. 13ff): The guidelines in general and especially this section should distinguish between professional AI systems (i.e., as used by business for business and public institutions) and retail-based consumer AI systems. The ethical boundary conditions or framework may be different as well as the road to making AI trustworthy. There is a big difference in realizing trustworthy AI for a professional / expert user (e.g., pilot, robotics operator, flight controller, etc.) or a lay person using a smartphone app or public institution AI-based app (e.g., tax declaration, social security, etc.).- Data Governance 1/2 (p. 14f): The statement that "the datasets gathered inevitably contain biases, and one has to be able to prune these away before engaging in training" (p. 14) depends on the aim of a given policy or algorithmic model. Pruning a model to make it fairer for one group may inevitably create biases and unfairness for another group, in particular if different groups have different descriptive distributions and base rates. It thus makes more sense to identify bias / unfairness with data that reflects the real world and then correct post-processing for bias and unfairness (which often would be relevant for minority groups in a machine learning setting).- Data Governance 2/2 (p. 14f): The following description does not appear to be correct: "It must particularly be ensured that anonymisation of the data is done in a way that enables the division of the data into sets to make sure that a certain data – for instance, images from same persons – do not end up into both the training and test sets, as this would disqualify the latter." (p. 15) Anonymization is not per se linked to which data is being used in train and test data, as long as the same data is not used in both sets. Two different pictures can easily be split and one ends up in train and the other in test data. This has little to do with anonymization. In fact, the process described may actually jeopardize proper train-test data splits. Furthermore, the section could benefit from a clearer description about legal data processing and legal grounds for processing of data from data subjects.- Design for all (p. 15): The statement that "systems should be designed in a way that allows all citizens to use the products or services" (p. 15) does not appear universally practical. It is highly likely that there will be AI-based products and services that appeal to particular groups rather than universally to all humans, e.g., gender specific apps, age specific apps (and combinations thereof). As rightfully stated later in the same section, AI applications cannot have a "one-size-fits-all approach".- Robustness – Resilience to Attack (p. 17): The requirements described in the guidelines regarding resilience and robustness apply to AI systems as well as to any ICT system (e.g. IoT system). Having said that, the guidelines would benefit from further considering the precautions that can be taken to raise the security level of AI systems. Highest security requirements should apply in AI development and application. All security features such as notification of security vulnerabilities, emergency stop button or security updates should be aimed towards a clear attribution of responsibility. Besides the risk of weak spots being exploited by hacker attacks, the self-learning capabilities of corrupted AI	- Ethics in autonomous systems and time-scales: We believe it to be instructive to look at ethics and intelligent autonomy based on time-scales. For example, do we allow AI systems to take full control and decision autonomy below a certain time limit beyond the human capability (e.g., below 500 milliseconds)? What situations do we allow this to happen in? What about minutes time scales in which humans start to be able to intervene in automated decisions or maybe do want to be in complete control? What about hours, days etc. where autonomous decisions may become reversible or iterative as more information becomes available? Note just because there is a lot of time to reverse a decision there may still be domains where we would not like AI to make such decisions without human oversight. However, below the 1-minute threshold there may be a lot of situations where it could be crucial to let the machine take control. This aspect has not been addressed at all but is very important. For example, network management functions such as beamforming in 5G networks – aimed at increasing spectrum efficiency – will require autonomous systems to make decisions in fractions of a second in order to ensure uninterrupted connectivity.- Consistent use of modal verbs: For consistency and to underline their guidance character, we suggest using "should" throughout the guidelines and avoiding other modal verbs such as "must". Mixing the use of modal verbs leads to inconsistencies. For instance, while it is made clear in the document that the guidelines are legally non-binding and can be voluntarily observed (or not), some of the language used contrasts that description (e.g. "data collected […] must be done in a way to avoid discrimination, manipulation or negative profiling" (p. 9); "this section lists five principles and correlated values that must be observed to ensure that AI is developed in a human-centric manner" (p. 8)).

		issues such as algorithmic pricing.	decision making capabilities and is not constrained by the use of an AI system.") it is important to ensure that AI does not constrain the private autonomy, i.e., individual freedom, of employees – or any human being for that matter. This should, however, not be interpreted as an individual right to object to any AI implementation in the working environment. For many professional applications AI is already today an inherent part of the working environment (e.g. pilots are supported by AI in aircrafts). Employees should participate collectively in decisions around the implementation of AI systems in working environments through established bodies of representation.- Principle of Justice (p. 10): Instead of stressing "that AI systems must provide users with effective redress if harm occurs" (p. 10), the guidelines should emphasize that ultimately humans are responsible. Operators of AI should know and make clear who is responsible for which AI system or feature.- Identification without Consent (p. 11): The draft guidelines raise concern about the "usage of anonymous personal data that can be re-personalized" (p 11). We note that the potential for re-identification highly depends on the technical means used to anonymize or pseudonymize data as well as the way data is being clustered, packaged and used after anonymization/pseudonymization. In the context of AI development and deployment, it is important to make more data available more easily, but as a principle, this should be done in a privacy-conscientious way.- Potential longer-term concerns (p. 12f): Artificial Moral Agents (AMAs) should not per se pose a threat as long as these have been trained within a given and acceptable ethical framework. It is highly likely that AMAs being trained by re-enforcement principles (where the reward is adherence to the ethical principles) are near-future feasible (i.e., white swans). This is decidedly not a negative development and might be one of the few technology principles existing today that might actually work in terms of developing practical ethical AI.	systems can exponentiate the risk of damage. For security in the development and in the application of AI, this specifically means: Ensuring IT security is a key requirement for product safety of AI applications or products that implement AI applications. This correlation must always be considered by developers and industrial users (security-by-design). Mandatory risk assessments (analogous to the Data Protection Impact Assessment of the GDPR) could contribute to highly sensitive AI applications such as in health-care. The current regulatory focus on operators of critical IT infrastructures, like in the ICT, health-care or energy sectors, is no longer sufficient, because referring to IoT, AI-based IT systems are being used increasingly, in which the criticality arises on an ad-hoc basis. This would for example be the case in a multitude of connected, self-driving vehicles.- Technical methods – Architectures for Trustworthy AI (p. 19): Trustworthy AI should not only be ensured by "formulating rules, which control the behaviour of an intelligent agent, or as behaviour boundaries that must not be trespassed" (p. 19), but also through mechanisms enabling operators to deactivate and stop AI systems at any time.- Technical methods – Traceability & Auditability (p. 20): We see a need to explain and clarify what specifically is meant with traceability and auditability – neither the guidelines' text nor glossary provide clear definitions. The guidelines should also better motivate the purpose for traceability and auditability since both are not ends in themselves. Moreover, any guidance in this respect should be context-sensitive and consider the purpose of an AI system, thereby differentiating e.g. professional or retail-based consumer AI systems (see above). Furthermore, explainability will also be enhanced if inappropriate data is removed to avoid bias. Operators of AI should keep track of decisions made and the information fed to the system in order to enhance decision quality.- Non-technical methods – Codes of Conduct (p. 22): The headline should rather read 'Internal Governance' as there is more to ensuring organisations adhere to ethical principles than codes of conduct. Furthermore, the section should reflect this. For instance, the scope of codes of conduct is the individual employee. Yet, not every employee has to deal with AI ethics. Therefore, DT has designed self-binding principles as a framework with complementary profession ethics etc. to ensure ethical AI development, implementation and use. - Additional technical methods: The guidance so far does not address appropriate safeguards given under the GDPR, that enable data processing in a privacy-friendly way. Data starts generating value when a significant amount is being processed, e.g. the "critical mass" of data can be reached – which is a key requirement for Artificial Intelligence. Pseudonymisation has an advantage vis-à-vis anonymous data, namely that the necessary "identifiers" remain intact for big data applications, to be able to merge large amounts of data from various sources. The technique thereby eliminates the direct link between data and data subject, while the pseudonym used as an identifier allows to repeatedly merge personal data from different sources over a period of time, which is a key requirement for valuable	

				data-driven services. It is also for that reason that the European Commission has embraced Pseudonymisation as a privacy-friendly technique in the GDPR, “to reap the benefits of big data innovation while protecting privacy”. (http://europa.eu/rapid/press-release_IP-15-6321_en.htm). The benefits of technical safeguards such as Pseudonymisation are thus inherent, more and more becoming the essential tool for companies to contribute to the data economy while at the same time appropriately protecting the interests of individuals.- Responsibility: The guidelines lack proper consideration of the need to responsibility in AI development, deployment and every-day use. Every AI system requires a clearly responsible person/role who/which decides in general about the operations of the system and has the responsibility to monitor its correct operation, i.e. a match between the things the AI system should intentionally do and the results of the things it really does.- Intervenability: In combination with transparency and responsibility among all other principles in this Chapter, intervenability is the mechanism which allows a person (privacy law: data subject) to object to the results of an AI system processing his/her data. He/she needs a defined instance where questions and complaints about the processing and the results can be placed.	
Alastair	McCapra	Chartered Institute of Public Relations			As the world’s only Chartered body for public relations professionals, members of the Chartered Institute of Public Relations (CIPR) have a responsibility to advise organisations and services on building AI technologies that will enhance human lives. Our Artificial Intelligence in Public Relations (#AIinPR) Panel has been actively encouraging our members and the public relations, communications and marketing industries across the globe to really understand how AI tools work and how their ‘decisions’ can impact human lives. The CIPR has a robust and globally-recognised Code of Ethics our PR practitioners must follow. As part of our leading #AIinPR panel work, we are encouraging our members to adopt a Code of AI Ethics in the organisations and businesses they work in and advise, whether it’s about data privacy, personalisation or deep learning. Ultimately, if we want to realise AI’s incredible potential in our public relations roles and also as a society, we must advance AI in a way which increases the public’s confidence that AI benefits society and upholds their trust in AI. Therefore, we believe there is a responsibility on all public relations, communications and marketing professionals to take the lead in helping organisations and business address key ethical questions surrounding AI. The CIPR, therefore, supports the guidance being proposed by the European Commission as a framework for promoting what an ethics-driven approach to AI should look like. This includes;

							• exploring how to avoid biases in AI algorithms that can prejudice the way machines and platforms learn and behave and when to disclose the use of AI to consumers and the public; • how to address concerns about AI's effect on privacy and responding to fears about AI's impact on jobs and society; and • areas of PR and marketing where companies and business must ensure AI doesn't inadvertently apply biases. CIPR Artificial Intelligence in PR panel www.cipr.co.uk/ai
Nicholas	HODAC	IBM Europe	The success of AI will largely depend on the Trust its users will have in it. Trustworthy AI correctly has two comments: technical robustness and an ethical purpose. AI Ethics have become an important element of competitiveness; they complement each other instead of contradicting each other. The intro also makes it clear that a tailored-approach is needed respecting and reflecting the different AI use cases/applications. A few comments: - The doc is indeed the "beginning", but the doc needs to clarify the process through which the discussion will continue, the guidelines will be updated, etc. - The doc needs to have as objective to become THE reference point for national discussion on AI Ethics in the hope of preventing the national AI initiatives developing a new set of AI Ethics guidelines (Europe does not need more fragmentation)	- Support the ethical principles, but AI developers/designers need more guidance in case there is a conflict between certain principles. Currently the Doc states "internal or external experts should provide guidance", but that is "too light"....Who will take the responsibility? Should there be an hierarchy in the Ethical principles? - "operate transparently" - First requirement should be that users are informed that they are interacting with an AI system, that is the first requirement under transparency. In addition, informed consent is not the only legal basis for processing under GDPR. The Doc on various occasions only refers to "informed consent" and creates confusion therefore. (same comment applies to section on critical concerns raised by AI)	- The requirements list the right topics, but the list can easily be streamlined through the "merger" of some requirements: "data governance + respect for privacy", "design for all + non discrimination", "respect human autonomy + governance AI autonomy", "robustness + safety" - "data governance" focuses mainly on data quality, therefore suggest renaming it....also, bias is sometimes intended and positive (the Doc gives the impression all bias are bad) - "design for all" focuses on "accessibility"? or is the requirement that AI should indeed be designed in such a away that it can be used by ALL? - "robustness"...need more guidance on how "accuracy" will be decided. Who decides what level is good enough? Also, the "fall back plan" needs to be assessed on its feasibility in all AI applications; might not be wished or needed or possible.	In addition to the questions that AI developers need to ask themselves we also need suggested responses. In some situations these answers depend on the AI applications/use cases and therefore providing suggested responses can be challenging; but some of the requirements/questions apply to all AI applications/use cases and therefore a generic response can already be provided.	

Anonymous	Anonymous	Anonymous	orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in

reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et

dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non

proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut

aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.orem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Roeland	Van der Stappen	Visa	Visa welcomes the opportunity to provide comments on the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) Draft Ethics Guidelines for Trustworthy AI. We are supportive of the concept of AI Ethics Guidelines as a flexible, evolving voluntary code of practice for industry and other stakeholders deploying AI, machine learning and automated decision-making algorithms either now or in the future. We also welcome work on a global approach to AI ethics, including the declaration on ethics and data protection in AI by the International Conference of Data Protection and Privacy Commissioners (ICDPPC).We see tremendous promise for AI in financial services and payments, while being compliant with the principles of the EU General Data Protection Regulation (GDPR). Indeed, if we are able to harness the power of data and AI in a way that upholds public trust, Visa believes the benefits could be no less than economically and societally transformational. In payments, fraud detection is currently the largest use case for AI, and the most significant application for Visa. Visa has used AI technology to augment proprietary Visa technologies to reduce fraud and enforce cybersecurity for over a decade. As a result of AI deployment, we can now identify and analyze fraud in near real-time using proprietary algorithms that process transaction data elements, while reviewing patterns of likely fraud for preventing attacks.We are now exploring an extremely broad variety of applications for AI beyond security and fraud detection. For example, Visa is partnering with third party companies to develop chatbot applications for Visa clients to use with their customers. These capabilities are integrated into the Visa Developer Platform. Our goal is to harness the power of data and AI for the benefit of customers, the businesses which serve them, and the global economy. We live in an increasingly digital world in which the ways data is created, stored and used are growing exponentially. As data generation and use explodes, there is growing concern about how technology companies are using that data. In order to fully realize the benefits of emerging technologies such as AI, consumers need to be able to trust that companies are using new technologies and consumer data ethically and responsibly. For Visa, trust has been the foundation of the Visa brand for over 60 years.Building and maintaining trust can be done in three key ways. Firstly, by individual organizations striving demonstrably to operate in a manner that continuously warrants the trust of consumers and regulators. This includes taking a pro-active, responsible approach to embedding principles and practices which promote public confidence in them. This should be a given for any organization with the power to impact people's lives, however it has not always been the case in the technology sector, as several high-profile examples have shown. Secondly, through industry collaboration and voluntary codes, such as the process being undertaken by the AI HLEG. Visa supports this approach as an effective way to promote coordinated good practice and risk management, without disincentivizing businesses from investing, competing and innovating. We agree with the AI HLEG that an ethical approach taken by industry will support responsible	Ethical PurposeIt is important for the designers of AI systems to be clear about the purpose of the system in order to build user trust and good governance. However, though the process by which the system was designed should be ethical, and a strong consideration of ethics should form part of the analysis of potential outcomes, it would be confusing for both designers and consumers if the stated purpose of every AI system was couched in terms of ethics. 'Purpose' is not the same as a statement of assurance that the designer does not intend to breach fundamental rights. Rather than conflating 'purpose' and 'ethics' in this way, it is preferable to suggest that designers have a clear statement of purpose as part of the principles around transparency and explicability; and that a separate assurance should be provided around the intention to design the AI system in line with the ethical guidelines which the organization has pledged to follow. Principles of Non-Maleficence and JusticeIt is inarguable that AI systems should not be designed with the primary purpose of harming individuals or society, whether such harms are physical, psychological, financial or social. Because the concept of 'harm' is perceived through the personal lens of the individual affected and does not have one singular definition, this principle as it is stated currently is so broad as to be unworkable. The potential harms to the individual should be carefully balanced against the positive benefits of the technology for society. For example, not harming an individual could indeed be impractical in the case of fraud detection, where the fraudster is going to jail but the societal benefits are clear. For this reason, the guidelines should maintain flexibility so that industry users can continue to use AI technology to support critical business functions, such as fraud detection and conducting valid risk-based authentication of payments.As the guidelines acknowledge, it is known that humans are biased in their decision-making. Since AI systems are designed by humans, it is possible that humans inject their bias into them, even in an unintended way. We must constantly be alert for outcomes, whether deliberate or negligent, which embed discrimination or injustice into the system; and in some systems we must design in bias to normalize results and remove unfairness (legitimate bias). The desired outcome is that no individual or group is unfairly prejudiced or excluded, and it is therefore more helpful perhaps to speak of 'unfair' or 'unwarranted bias'. A useful, though theoretical, comparison would be whether the level of bias would have been greater or lesser should purely human decision-making have been used.	Transparency & explicabilityAs a general rule, we believe a willingness and ability to be honest and open with consumers and disclosure regarding AI usage in products and services will enhance users' trust in the technology, facilitate uptake and protect firms' license to operate. However, whilst supporting the concept of a trustworthy AI, one should be working towards the requirements not being too stringent and ensuring they do not hinder companies fulfilling other regulatory obligations, such as prevention and detection of financial crime, terrorist financing and cybersecurity incidents.In that regard, we would support exempting the use of AI to prevent and detect financial crime, terrorism financing and cyber security incidents from a transparency and explicability obligation. Individuals should not be aware of how the technology works in these cases, as it would risk underlining the purpose of the anti-fraud detection systems. However, transparency towards financial supervisors and law enforcement may be warranted. Similarly, we would also not wish to see genuinely beneficial, and ethical, uses of AI restricted because it is too complex, or too commercially sensitive, for organisations to be fully transparent about the design or operation of the system. If firms fear being forced to disclose or share valuable IP, the inevitable results will be a reduction in innovation and a move to markets with more flexible conditions. In 'hard to explain' or 'too sensitive to explain' cases, again there may be a spectrum of transparency and explicability, based on the level of information provided to different parties. If the purpose is to build trust and consent, a statement of purpose, combined with an assurance of ethics and the ability to challenge and rectify outcomes, may be sufficient in those instances where only limited transparency and explicability is possible or desirable.		Align legal basis for AI data processing with GDPR To date, payment networks such as Visa are able to process personal data for 'legitimate interests', including for the purpose of fraud detection and prevention, under Art. 6 of the GDPR. However, the draft AI Ethics Guidelines refer to 'consent' as the sole basis for AI data processing, which is more restrictive than the GDPR and hence would inhibit our fraud detection and prevention systems. We suggest the legal basis for AI data processing aligns fully with the GDPR. Individuals would still receive a suitable description of the data collected, how it is processed, and with which third parties it will be shared. The latter is required by the GDPR regardless of the legal basis for data processing, albeit 'legitimate interests' or 'consent'.

			competitiveness and the development of cutting-edge, secure AI in Europe.The third element, an effective regulatory environment, will also be important. Regulation and policymaking in the digital age presents new and challenging problems. The shift toward intangibles, such as AI-based products and services, represents a significant macro-economic development. Thus far, even the most sophisticated regulatory approaches have not kept pace with the speed and nature of technology innovation, and the resulting impacts on society. Many of the traditional consumer outcomes which regulation aims to encourage (price, quantity, quality, innovation, choice) are being re-evaluated in the context of additional concepts which may be as, or more, appropriate in the digital age, for example fairness, consent, privacy and democracy. An optimal regulatory environment is one that is not overly prescriptive while encouraging innovation and protecting end users; and is likely to result from a thoughtful, collaborative long-term effort between industry and regulators, rather than a definitive set of rules as may have been the approach in the past. As AI technology is still evolving, we prefer fair market standards to regulatory standards. Furthermore, stable market standards can only be identified once there is sufficient innovation in the market, to ensure that innovation is not hindered		
Kai	PETERS	VDMA - Mechanical Engineering Industry Association	VDMA, the German Engineering Association, welcomes the work of the High-Level Expert Group on the "Ethics Guidelines for Thrustworthy AI" as an important contribution to the debate on ethical and societal aspects of digital technologies. In order to harness the societal and economic benefits of AI, it is essential to ensure its lawful and ethical use, acceptance and trust. The focus of the ethics debate and the the guidelines should be more clearly on applications where human dignity, human rights and core values of society are at stake. These questions are not equally relevant for every application scenario: Artificial Intelligence is a collective term which describes a wide range of technologies, products and services which could be used in many possible applications, showing a different degree of ethical criticality. Therefore, on the one hand, the guidelines must focus on ethical and political challenges. On the other hand, the guidelines must also acknowledge that there is a vast range of application scenarios which are ethically less critical - because the problem is purely technological, no humans are involved or societal impact is very limited. In addition, more consideration should be given to the fact that often potential AI-issues are already subject of EU-law and existing rules. It is essential that guidelines are consistent with existing law and that no parallel, technology-specific frameworks are established. This lack of differentiation could lead to drawing too generic red lines, which might unnecessarily limit the scope for innovative and valuable AI-applications. For instance, the formulation of generic horizontal "critical concerns" in chapter II.5 might lead to very		In general, we suggest to take up a more positive and moderate point of view by acknowledging that there are areas with low risk and low criticality, but with high value and high impact. In these areas, innovation must not be stifled by non-relevant rules, but should be encouraged by concepts such as the "innovation principle" or "curiosity-driven development". It should be also taken into account that there are societal costs of not using technologies. For the European engineering industry, AI-technologies are opportunities to maintain worldwide leadership, to increase competitiveness and sustainability. Applications of industrial AI such as predictive maintenance or optimization of the use of resources show that the use of AI in industrial applications offers many opportunities and promises considerable benefits. VDMA is looking forward to a fruitful continuation of the debate and is open for questions or comments with regard to these aspects or other questions related to the use of AI-technologies in industry.

			rigorous and undifferentiated requirements. VDMA disagrees with parts of the definition proposed in the draft document. The use of wording such as “perceiving” and “reasoning” in this context is misleading and raises the image of AI-systems which equate or resemble humans. Furthermore, the inclusion of decision-making as an inherent characteristic of AI could lead to wrong conclusions about the general autonomy of AI-systems. Therefore, the VDMA proposes to amend the definition as follows: Artificial intelligence (AI) is a collective term and refers to systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and providing predictions or results, which might be implemented automatically, if considered appropriate and desirable."				
Pascal	GAREL	European Hospital and Healthcare Federation	The rationale is clearly presented and the overall logic is understandable. However even if the Guidelines present themselves as not being a "substitute", it will be difficult to avoid misunderstanding between them and the policy making/regulation. The next deliverable will certainly be interesting to read. More generally jumping directly to Guidelines created by a group of "experts" without a proper political debate seems to confirm those who expresse theirs doubts about the legitimacy of the European union consultation procedures.	As far as health rights (as presented in the EU Treaties and the Charter) are concerned, they have not been enough developed and it is not enough to include them later in the use case.		The page 28 mentions a use case on healthcare diagnose and treatment and ask for thoughts on how the assessment list should be construed for and applied: we suggest to ask DG SANTE as a natural partner for all of us involved in the healthcare to lead a special work, not limited to an on-line consultation.	
Anonymous	Anonymous	Anonymous	correct and acceptable	correct and acceptable	correct and acceptable	correct and acceptable	correct and acceptable
Bart	Verheij	CLAIRE Informal Advisory Group on Ethics, Legal, Social Issues (CLAIRE IAG ELS, claire-ai.org/IAGs/#ELS)	As the CLAIRE Informal Advisory Group on Ethical, Legal and Societal Issues (CLAIRE IAG ELS), we applaud the discussion of ethics guidelines for trustworthy AI building on European institutions, such as the EU Charter of Fundamental Rights. As an aim of the document is to stimulate a thriving European AI ecosystem with a recognizable European flavor, it seems wise to invest in innovative AI techniques with already strong basis in Europe. In particular, we would like to emphasise the following technical methods that can support an ethical, legal and social AI, concretely:- Computational critical thinking and argumentation- Combined logical and learning AI methods (with on the one hand knowledge representation and SAT solvers, and on the other Bayesian networks and deep learning)- Rule-based and case-based reasoning models- Cognitive and neuromorphic system designs building on research in cognitive and social science (theory of mind, heuristic reasoning, cognitive materials)These and related technical methods are for instance investigated at technical conference series such as COMMA (on argumentation systems, since 2006), DEON (on normative systems, since 1991), JURIX (on legal AI, since 1988), ECAI (on AI, since 1974), all of them founded in Europe.We would also like to	[Repeated from comments on the introduction]We would also like to invite the EU HLEG AI to provide more clarity on the relationship between the Ethics Guidelines and existing laws and regulations. The assumption (on p.3, top) that compliance with the Guidelines can only complement legal and regulatory compliance seems too simple, as the relationship between ethics and law can work in both ways (i.e. ethics not only complement existing laws and guide their interpretation, but may also challenge existing laws and inform new rules – cf. the notions of soft versus hard ethics used by Floridi in Phil. Trans. R.Soc. A 376: 20180081; http://dx.doi.org/10.1098/rsta.2018.0081). It is also recommended that the Ethics Guidelines at least explicitly acknowledge that trade-offs between conflicting rights, principles and values will have to be made, and, in a more ambitious scenario, provide further hints on how to deal with such cases of competing interests (eg. what are the criteria to strike a fair balance? How to organize this balancing exercise within the company?).	[Repeated from comments on the introduction]As an aim of the document is to stimulate a thriving European AI ecosystem with a recognizable European flavor, it seems wise to invest in innovative AI techniques with already strong basis in Europe. In particular, we would like to emphasise the following technical methods that can support an ethical, legal and social AI, concretely:- Computational critical thinking and argumentation- Combined logical and learning AI methods (with on the one hand knowledge representation and SAT solvers, and on the other Bayesian networks and deep learning)- Rule-based and case-based reasoning models- Cognitive and neuromorphic system designs building on research in cognitive and social science (theory of mind, heuristic reasoning, cognitive materials)These and related technical methods are for instance investigated at technical conference series such as COMMA (on argumentation systems, since 2006), DEON (on normative systems, since 1991), JURIX (on legal AI, since 1988), ECAI (on AI, since 1974), all of them founded in Europe.		[Repeated from comments on the introduction]As the CLAIRE Informal Advisory Group on Ethical, Legal and Societal Issues (CLAIRE IAG ELS), we applaud the discussion of ethics guidelines for trustworthy AI building on European institutions, such as the EU Charter of Fundamental Rights.

			invite the EU HLEG AI to provide more clarity on the relationship between the Ethics Guidelines and existing laws and regulations. The assumption (on p.3, top) that compliance with the Guidelines can only complement legal and regulatory compliance seems too simple, as the relationship between ethics and law can work in both ways (i.e. ethics not only complement existing laws and guide their interpretation, but may also challenge existing laws and inform new rules – cf. the notions of soft versus hard ethics used by Floridi in Phil. Trans. R.Soc. A 376: 20180081; http://dx.doi.org/10.1098/rsta.2018.0081). It is also recommended that the Ethics Guidelines at least explicitly acknowledge that trade-offs between conflicting rights, principles and values will have to be made, and, in a more ambitious scenario, provide further hints on how to deal with such cases of competing interests (eg. what are the criteria to strike a fair balance? How to organize this balancing exercise within the company?).On behalf of the CLAIRE Informal Advisory Group on Ethics, Legal, Social Issues (CLAIRE IAG ELS, claire-ai.org/IAGs/#ELS)Disclosure: some members of the CLAIRE IAG ELS are also members of the EU HLEG AI.			
Wojciech	JAMROGA	Institute of Computer Science, Polish Academy of Science			Section 2.1 (Technical methods): I would suggest to involve formal methods in a more systematic way. - On one hand, we should attempt *formal specification* of the requirements mentioned in Section 1. - On the other hand, the methods for testing & validating (p. 20) can be augmented by formal verification (especially model checking) and simulation. - It is clear that formal specification can only approximate the intuitive requirement. Similarly, formal verification can never give full confidence in the correctness of the verified system. Still, they can help uncover the internal structure of informal requirements, and detect potential vulnerabilities. This fits very well the circular model of development in Figure 3.	See above. Formal specification and verification can and should be a part of both development and assessment of Trustworthy AI.	
Dhananjoy	Mishra	AI ETHICS INTERNATIONAL	Scope of the Guidelines- "..... not legally binding". Why not binding? This is keeping an window for deploying cheaper versions and thereby diluting the effort and intention. B.(II)- Realisation of trustworthy AI- reporting to authority , incase of some non-acceptable activities found, should also be put as a responsibility for developers and users. This will help in gathering more information and also addressing issues at the starting stage. Framework of Trustworthy AI- Fig-1 We can add up a procedure of clinical test, like what we do in case of pharma drugs.	5.4- LAWS - can we escalate level or initiate the requirement that there should be a mechanism similar to 'Non-Proliferation Treaty' for nuclear weapon.	3. Design for ALL- 'that allows all citizens……., regardless of age,…..' Why so? There could be specific AI meant for specific use by certain types of citizens, even some cases by other living animals. Or in future by AI itself. We don't allow children to drive car , isn't it? Similar issues may also required to be addressed. 8. Robustness – Is it possible that input data are classified and other situations are disabled? Initially it will restrict growth but will ensure safety. 10. Transparency – " Being explicit and open ….." Transparency will have a limitation. We cannot be so transparent on development that our IP is diluted and copied by competitors. 1. Technical Methods Architectures for Trustworthy AI- Can we work on a system, that the impact of the architecture is assessed in low risk system and once qualified , then only upgraded to a higher system. In future, there will be AI modules available in the market , which can	Assessing Trustworthy AI- 2nd paragraph ' the primary target …….indirectly with humans….' Why not other animals and Mother Earth also included. We have already done enough harm to them . 1 Accountability- External Auditing- Can we have a rating for this auditors? Or certain requirement on specific expertise domain? In future, there will be AI modules available in the market , which can be used standalone or can be integrated to a complex AI . Like we presently purchase the chips and integrate them to a circuit. Who will be accountable to such situations. There should be an accountability matrix for both user and OEM. 3. Design for ALL-Is it a system that can be used for or by children? 4. Governing AI Autonomy 'Within the organization who is responsible ……' How he/she qualifies for such position? What is the methodology or minimum criteria? Reliability & Reproducibility : In case of	Glossary- AI difinition has considered 'systems designed by humans'- What about a situation, when AI system will be developed by machines or AI. Any of the guidelines or stipulated conditioned of this guidelines can not be enforced in such situation. Like the Law on Fundamental Rights is binding, Ethical behaviour needs to be binding. Otherwise it will be diluted and the ultimate goal will not be acheived.

					be used standalone or can be integrated to a complex AI . This components thus will go through a qualifying procedure, before being tested/used in complex models. Testing and Validating- Can we include some 'proof tests' or 'stress tests'; criteria for which can be included in the next level document? Traceability & Auditability- " to tackle ……mistakes " How much will it be practical for a 'deep learning' situation, where a significant layers are involved? And thus, this requirement shouldn't hold the innovations or technological growth. 2. Non-Technical Methods Regulations : Humans and other animals are also not above law. Incase of a breach that is not acceptable, the USER and the OEM can be booked. More importantly, depending on the severity , the system the generated that particular AI, can be barred, like we put an offender behind bar. Codes of Conduct : There should be a mention of a separate set of KPI on Ethics parameters . The outcome of the tracking on this parameter will help growing the trustworthiness of the system.	standalone modules are available , which are integrated by another integrator to a complex AI, who remains responsible to ensure the reproducible performance ?	
Dominique	BONETTI	Grande Loge Féminine de France, Institut Maçonnique Européen	Principales lignes directrices pour garantir un objectif éthique• Veiller à ce que l'IA soit centrée sur l'humain : l'IA devrait être développée, déployée et utilisée dans un « but éthique », fondé sur les droits fondamentaux, les valeurs sociales et les principes éthiques de bienfaisance (faire le bien), de non-malfaisance (ne pas nuire), autonomie des humains, justice et explicabilité. Ceci est crucial pour aller vers une IA digne de confiance.• S'appuyer sur les droits fondamentaux, les principes éthiques et les valeurs pour évaluer de manière prospective les effets possibles de l'IA sur l'homme et le bien commun. Porter une attention particulière aux situations impliquant des groupes plus vulnérables tels que les enfants, les personnes handicapées ou des minorités, ou aux situations d'asymétrie de pouvoir ou d'information, telles qu'entre employeurs et employés, ou entre entreprises et consommateurs.• Reconnaître et être conscient du fait que, tout en apportant des avantages substantiels aux individus et à la société, l'IA peut également avoir un impact négatif. Rester vigilant pour les domaines critiques.Conseils clés pour réaliser une IA digne de confiance• Intégrer les exigences relatives à l'IA de confiance dès la première phase de conception : responsabilité, gouvernance des données, conception pour tous, gouvernance de l'autonomie de l'IA (surveillance humaine), non-discrimination, respect de l'autonomie humaine, respect de la vie privée, robustesse, sécurité, transparence.• Envisager des méthodes techniques et non techniques pour assurer la mise en œuvre de ces exigences dans le système d'IA. De plus, tenir compte de ces exigences lors de la constitution d'une équipe chargée de travailler sur le système, le système lui-même, l'environnement de test et les applications potentielles du système.	Point clé : éduquer à mieux connaitre et appréhender l'Intelligence Artificielle en tant qu'outil en évitant les qualifications ambiguës Dans le draft, les termes « raisonner » et « raisonnement » sont utilisés pour désigner le mode de fonctionnement de l'IA. Or, l'on doit se poser la question de différencier les processus de l'humain et les évolutions technologiques, il faudrait pour cela éviter d'utiliser le mot « intelligence » pour des algorithmes.- L'usage du terme « raisonnement » sous-tendrait l'avènement d'une « IA forte » capable de copier le fonctionnement du cerveau humain. Or à ce jour, le cerveau humain et les processus cognitifs ne sont pas modélisable et il n'existe aucune théorie proche d'offrir un modèle même approché – soit du cerveau, soit de la « rationalité ». - L'Intelligence artificielle dépasse déjà nos aptitudes dans notre capacité à calculer, mémoriser ou discerner des détails. En outre, l'IA peut déjà être considérée d'égale à égale sur une traduction, un raisonnement spécialisé, voire une détection d'émotions.- Dans la pensée il y a une conscience, dans un logiciel informatique il y a une action et pas de conscience. La déduction n'implique pas une conscience. La notion de finitude est très importante dans le développement de la conscience humaine tandis que la notion d'infinitude est induite dans le programme informatique. - Notre groupe de travail reste divisé en ce qui concerne la nature d'une IA en devenir : nouvelle forme de conscience, fruit d'une complexification de la matière résultante de la densification des connections dans les ordinateurs quantiques, ou évolution technologique exponentielle associée à une cognition augmentée. - Nous aimerions donc souligner l'importance de la terminologie employée pour désigner n ensemble de technologies et recommander en particulier de cesser d'employer le terme d'Intelligence artificielle.- La question sur laquelle nous sommes toutes d'accord : c'est comment la contrôler et rester vigilantes ?	Points clés : L'éthique de l'Intelligence artificielle doit faire partie du dispositif de gestion des risques de la responsabilité du conseil d'administration (risques sociétaux)Il est de la responsabilité sociale de l'entreprise de donner au citoyen les outils de compréhension et de préservation de la liberté de conscience face aux IA qu'elle déploieLe document rédigé par le HLEG est très théorique et fixe le cadre de référence sans vraiment expliciter comment vérifier et mesurer le respect de ce cadre. La mesure de l'éthique est effectuée en regard des principes normatifs des droits de l'homme. Mais Le rapport fait peut-être l'impasse sur des sujets de bien collectif, de bien individuel, qui peuvent comporter une certaine relativité. Il existe un risque que la façon de définir la norme entraine une convergence et nuise à la diversité, qui peut également être considérée comme un bien. Il est complexe de mesurer les effets sociétaux induits par ces mécanismes à grande échelle.Nous pensons que pour adresser de façon concrète les problématiques éthiques, il faut examiner concrètement le rôle des acteurs qui sont les sociétés, les individus et les états ou organes de régulation.En ce qui concerne le point II, réalisation d'une IA digne de confiance, les responsabilités reposent essentiellement sur les sociétés qui mettent en œuvre l'IA. Donc, elles doivent mettre en place la gouvernance appropriée, en interne et avec leurs clients. Elles doivent rester à l'écoute des règles et des évolutions proposées par la société civile et par les tiers de confiance.A- Les sociétés qui utilisent l'IA sont confrontées à 2 types d'enjeux : l'utilisation de l'IA dans les produits et l'utilisation de l'IA dans l'organisation du travailL'IA va bouleverser les processus internes et l'organisation des entreprises : cela devrait se faire dans le respect des salariés et du sens donné au travail de chacun. En allant plus loin dans le découpage des tâches, l'IA créerait un néo-taylorisme déresponsabilisant ou au contraire permettrait aux personnes de s'épanouir par des organisations plus	Points clés : L'évaluation des risques liés à l'IA repose sur la vigilance du citoyen et des associations professionnelles.Il faut créer un régulateur indépendant garant du respect des principes éthiques, ayant accès à toutes les données et doté des moyens de contrôle et qui peut être saisi par les citoyens.A. Le citoyen, l'individu et les associations professionnelles sont au cœur de la détection des biais et des dérives. Ils doivent disposer d'un système de remontée des anomalies à un régulateur national. Ils doivent pouvoir agir en tant que collectif, formant par exemple des comités citoyens qui pourraient tester les IA avec une diversité de profil, ou en tant qu'associations professionnelles composées de personnes expertes, à même de comprendre les éventuelles dérives rapportées dans leurs domaines.B. Un organe de régulation des IA à l'image des régulateurs de la Banque et de l'Assurance pourrait traiter les alertes citoyennes : L'état / l'Europe a le pouvoir d'imposer des lois et de mettre en place des organes de régulation nationaux et supra nationaux.Nous avons la conviction qu'il faudrait mettre en place un système de régulation qui puisse agir en tant que tiers de confiance, ayant accès à toutes les données pour en analyser la qualité et les biais, qui puisse vérifier l'application des principes éthiques, comme dans le cas des banques pour les évaluations de risques systémiques. Le régulateur pourrait être saisi par tout citoyen qui identifie un biais et serait ainsi le centralisateur et le médiateur des plaintes relatives aux IA.La question des biais se pose différemment : nous savons que les données humaines sont naturellement biaisées. Il faut alors rectifier les algorithmes en fonction de critères choisis pour corriger les injustices induites par les biais humain. Qui fixe ces critères et comment est un sujet éminemment politique qui souligne bien l'importance du point 2 de formation des citoyens et la nécessité de les impliquer dans le processus. Par exemple :• Dans le futur, hommes et femmes devraient être sur un même pied d'égalité. Peut-on	L'Institut Maçonnique Européen de la Grande Loge Féminine de France espère que ces bouleversements et progrès scientifiques seront réalisés au bénéfice de l'Homme et que seront préservés dans les temps à venir la Liberté, l'Egalité et la Fraternité qui fondent notre société civile.

agiles.L'utilisation de l'IA pour la commercialisation ou dans les produits peut également avoir des impacts sociétaux.Aussi, les sociétés doivent prévoir des algorithmes construits, dès les phases de conception et de pré-lancement, pour respecter les principes éthiques (« Ethic by Design ») et qui sont constamment suivis et vérifiés.Le respect des principes s'accompagne de la mise en œuvre des mesures suivantes :1) Garantir la Précision des algorithmes : c'est l'analyse technique qui permet d'évaluer leur fiabilité, notamment le risque d'erreurs dans le système et le risque de préjudice pour les utilisateurs. Il est alors nécessaire de prévoir le processus de détection et correction des erreurs ; 2) Créer des outils permettant une compréhension suffisante des utilisateurs : explicabilité ;3) Mettre en place des Tiers de confiance pour vérifier les algorithmes sur la base d'un échantillonnage ou par des jeux de tests spécifiques : auditabilité ;4) Créer des normes relatives à l'impartialité : pour juger de l'absence de biais vis-à-vis de groupes ou de catégories de population, il faudrait inclure un algorithme d'exploration de données qui tient compte de l'équité. Mais la vision éthique d'une société comporte des éléments relatifs à la culture et à la période ;5) Introduire les risques liés à l'utilisation de l'IA, qu'ils concernent les changements d'organisation ou l'impact social des produits et des services dans la cartographie des risques ESG (Environnementaux, Sociétaux et de Gouvernance) pour la partie des indicateurs sociétaux. Définir une chaîne de responsabilité, comme pour la RGPD (Règlement Général sur la Protection des Données), pour donner des réponses rapides en cas de problème. Cette chaine de responsabilité doit remonter jusqu'au conseil d'administration via l'inclusion dans le rapport obligatoire sur la RSE (rapport de responsabilité sociale d'entreprise). Toutes ces mesures pourraient être contrôlées en interne par un responsable de l'éthique des algorithmes. Par rapport à leurs utilisateurs, les créateurs d'IA doivent mettre en place les conditions d'une utilisation éclairée de l'IA, par les moyens suivants :B- Le citoyen / utilisateur de l'IA : il doit disposer des outils de compréhension qui permettent de préserver sa liberté de conscience et d'identifier des biais. Ces outils sont :1) d'une part des interfaces qui rendent transparents, compréhensibles et auditables les facteurs qui ont conduits à la proposition, des interfaces qui permettent d'éviter la manipulation mentale en donnant un recul et une diversité d'offres;2) D'autre part la formation des citoyens : il n'en est pas fait état dans le document alors que c'est une des clés pour une utilisation éthique de l'IA. Elle pourrait être partiellement prise en charge par les sociétés qui déploient l'IA. Cette formation pourrait être complétée par une plateforme européenne adaptée à chaque état comprenant la formation par des MOOC, des cahiers de biais avérés…Il s'agit d'un enjeu majeur de politique publique3) Information obligatoire pour les services et produits sur la nature des algorithmes utilisés et les risques identifiés induits, notamment relatifs aux données personnelles utilisées et aux facteurs pouvant biaiser ou influencer le comportement, avec des exemples illustratifs à la portée de tous. Cette information est particulièrement cruciale envisager une politique pro-active pour rectifier ce biais du langage de façon opérative ? L'IA offre le moyen de rectifier les biais humains par l'éducation des algorithmes. Il faut alors prendre de la distance par rapport à une photographie de l'existant, par exemple en introduisant d'autres biais qui réintégreraient le féminin mais aussi des valeurs de notre société future. Ne pas sexuer les chatbots par ex. ; • Prévoir une brigade « anti-manipulation mentale » en charge d'identifier et prévenir l'usage de l'IA pour manipuler de groupes d'individus dans des buts criminels, terroristes, ou de privation de leur libre arbitre (ex. manipulation d'élections, recrutement par des réseaux terroristes) ;• Imposer une procédure de « débranchement » des algorithmes et imaginer comment débrancher les algorithmes de surveillance, par exemple chinois qui ne disposeraient pas de ces normes• Aborder le sujet en termes de biais souhaités et apporter des éléments de quantification

					dans le cas d'IA embarquées dans des Robots Humanoïdes.	
Anonymous	Anonymous	Anonymous	First I'd like to say how I admire the work you've done. Considering a collective intelligence as an artificial one, since it is human made, you're guidelines could be of great inspiration for a society looking to increase sustainably based on its knowledge without risking to destroy itself. So, I love that Europe is taking the initiative to write them, as I see it as some new Siècle des Lumières, where european thinkers were trying to reconsider humanity as a whole. Considering your work as the product of an artificial intelligence, you should apply your guidelines to the development of your project (maybe you already are) itself, meaning you should ensure diversity and thus work with experts from other continents to make sure we don't make rules that carry our bias. Even if it goes against the goal, "which will enable Europe to become a globally leading innovator in ethical, secure and cutting-edge AI", since all the countries taking part would share the method. This last goal is not human centric enough, and the project has to be, if you want it to be ethic. It should enhance the well-being of all Earth inhabitants (I am pretty sure that if you work with an Hinduist, he'll try to develop a Cow Centric AI. It's a joke. I've read in your document that "the various life forms should be protected"). After Fig.2 (Chap. I.2) : "In 1997, the members of the Council of Europe adopted an instrument called the "Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine"" Do you know if Europe tried to work closely with the chinese on this matter ? Recent news about gene-edited babies could let us think they did not engage in such ways. Any parties taking part in the project should		p.24 (Accountability), you write : "Was a diversity and inclusiveness policy considered in relation to recruitment and retention of staff working on AI to ensure diversity of background?" Do you have a guy with my background ?	

			though apply the five core ethical principles of said guidelines, including explicability as a vow of transparency, which means even private parties could join the project - if governments could not agree on those terms - implying they'd have to share the results of their R&D Department with the group (open source baby! Or "public sources") and maybe pay a tax to the public entities they would be working with (maybe as a part of producted AIs, for the public services of all the countries engaged in the process). I am just giving spontaneous, and maybe ridiculous, ideas to explain that there may be models that could permit to maintain a human-centric goal, without risking to completely lose EU market shares to a more resourceful foreign producer. EU could have parts in said products and the producer could be labeled "life centric"				
Juan Ignacio	Rouyet	We The Humans	No comments	No comments	CHAPTER. 1. No comments. In the ten requirements, we would like the document to include the limits and parameters per each use case and per IA technology type (chatbots, machine learning, autonomous vehicle, robots, and etcetera.) We consider not every requirement applies in the same way to the different use cases and technologies	Beyond the questions in the assessment list, we would like to have concrete mechanisms for the fulfillment of the requirements in different situations or cases. One of this mechanism has to do with the roles in charge of control and management.	- One thing to be determined and specified is the area of application of this ethical guide. Who should know it and follow it and where? We would say this is an Ethical Guide with European Context and Global Application. - We want to suggest that in further developments there should be mechanisms of ethical certification and punishment in case of a regulatory breach. - We would suggest getting more concrete tips for helping these ethical implementations in the companies, similar to those already in place for the Quality Certification Systems. - We would suggest a definition of the different stakeholders: developers, end users, IA providers, investors, companies which will use IA, and etcetera.
Anonymous	Anonymous	Anonymous					
Jacob	Dexe	RISE - Research Institutes of Sweden		5.2 Covert AI system: There are multiple other examples of AI systems that can be plausibly affected by this writing. Perhaps to the extent that is not a matter of just border-cases and should therefore warrant a less severe start of the paragraph. When a person talks to a human aided by an AI system - is that supposed to be acknowledged? When the application is limited in scope and intelligence (e.g. getting a sales quote generated by an algorithm) - is that supposed to be acknowledged?	1.2. Data governance: "The datasets gathered inevitably contain biases, and one has to be able to prune these away before engaging in training" - The thing that is important is not being able to prune awa biases, but being able to identify them and make informed choices about how to deal with them. Pruning away is one action to be taken when you've identified the biases that are in the datasets. Also, training might be needed in order to identify the biases, be it a testrun or a continuous update of the algorithm in order to deal with new weights or biases. This point is made in the "general comments" section as well, as it is important for the entire guideline. 1.4. Governance of AI autonomy: An added problem here might be that oversight of the quality of an AI system might become more difficult as fewer humans are performing the same task as the AI on a regular basis. Small inherent errors will be harder to adjust for in an algorithm than the larger random errors that humans might make. As human experience and routine disappears, either because they stop performing the task or because they become reliant on AI support, the importance of quality control ought to increase and regular independent benchmarking might become	While the list is comprehensive, it is also far from covering every relevant aspect, and as such it makes it hard to give additional comments on each section. It is important to consider that it is hard to thing outside of your own box. You can't readily imagine things that are outside the scope of your own imagination and competence. In large organisations it might be easy to have diverse teams that can work together on ethical guidelines and consequence analysis, but in smaller organisations that's not always possible. On the other hand, those organisations might not even get to these guidelines because of the same constraints.	Bias: While it is in line with the common understanding of the concept, there are reservations to be made on how "bias" is understood and used throughout the text. Especially when we try to understand what an absence of bias is. Is a bias alwayys something you want to remove from an algorithm? Is an algorithm completely void of biases a "better" algorithm than one that has biases in it? For many narrow uses of AI, a complete removal of bias is perhaps not the point one wants to reach, but instead we might want to understand what biases exists in the data collection, processing and optimization, and make informed decisions about which are acceptable or not. In other text the phrase "unwanted bias" have been used to highlight this point. It would be beneficial for the guidelines to adopt that phrase as well. A bias towards certain universities might not be an actual problem for a company making an exercise app that uses AI, and it might even be an active choice for that exercise app to have a bias towards non-smokers, or other moral positions. For a government or public authority, or for a supplier of a regulated service, the removal of biases might be more important, but it has to be a question of

				more important. 1.10 Transparency: A point well made. Perhaps it would be even better with additional emphasis on transparency in how the company works _with_ the AI systems, apart from how the AI systems works itself.2.2 Non-Technical Methods: While it is implicit in some of the points made it might be relevant to also highlight developing and applying policies on how organisations want AI systems to work and what limitations should apply.		context, as is pointed out under "Scope of guidelines" as well. Not least because the removal of all bias is a hard task to solve, and limiting it to "unwanted bias" might help both implementation of AI and the discussion about bias.
Lorenzo	Fiori	Fondazione Leonardo Civiltà delle Macchine	With reference to the rationale, structure and foresight of the Guidelines, Fondazione Leonardo Civiltà delle Macchine would like to share its observations and suggestions on the following issues: - An ethical AND legal framework: the research, development and use of Artificial Intelligence (AI)'s technologies and applications could be easily tested against a relatively "simple" matrix, which would take into consideration both ethical and legal factors. On the ethical side, AI use and applications should be considered in the light of good/right and evil/unfair, with specific reference to the universal values of peace, respect for human life and, no less important for the present and future of humanity, respect for the environment. On the legal side, the same technologies and applications should be assessed by taking into consideration what is lawful/permitted and what is unlawful/prohibited, always looking at fundamental human rights such as freedom, equality, democracy and privacy, the latter being of the utmost importance given the value of data for AI. While conciliating both the moral and the legal approach to AI, such complete framework would also provide a clear and comprehensible model and could represent, at the same time, a flexible tool to progressively evaluate the changing impact of AI on the medium-long term. - A balanced approach to Narrow and General AI: while the side document published by the HLEG (A Definition of AI: main definition and scientific disciplines, 18/12/2018) provides an acceptable and comprehensive definition of AI, the draft report itself does not include a clear distinction between Narrow AI and General AI. On the one hand, Narrow AI can handle single specific tasks but is already widely used in several daily applications (ranging from musical software to autonomous vehicles). On the other hand, General AI refers to an eventual future stage, when AI could achieve cognitive abilities and a general experiential understanding of operational environments akin to that of humans, being able to elaborate data and solve the most complex tasks at an infinitely higher speed. While General AI would eventually have more long lasting and disruptive implications than Narrow AI from the societal and ethical point of view, requiring new and more demanding ethical and regulatory standards, such breakthrough could only eventually occur in the long term. In the meanwhile, as a precautionary measure, it would be wise to broadly define what ethical requirements should apply to an eventual General AI.	With reference to this chapter, Fondazione Leonardo Civiltà delle Macchine would like to share its observations and suggestions on the following specific points: - Fundamental Rights of Human Beings (p.7): Among fundamental citizen rights, section 3.5 mentions the right of citizens to express opt out from any system of automatic treatment of their own data. Later on, section 4. on Ethical Principles in the context of AI (p.8), refers to the right of withdrawal from direct or indirect AI decision making. These categories of rights protect the autonomy of the individual only ex post, after he or she has been subjected to the screening of AI based algorithms. The framework of the document could benefit from adding the concept of the "right to opacity": the individual right to remain invisible to the screening and control of any AI automated system, without prior consent. This would reinforce the Principle of Autonomy (p.9) and allow protecting human agency also a priori. We are however conscious that it might prove difficult to implement the right to opacity from both the technical and political point of view. On the one hand, the right to opacity might force AI service providers to always keep humans in the loop of automated decision-making, allowing human control over the initial screening of users but also diminishing the productivity gains allowed by AI. On the other hand, AI providers might decide to restrict the range of their services only to those users that accept the automated treatment of their data. This might engender problems of political and social exclusion, particularly when AI is applied to the state domain. - The principle of Explicability: "Operate Transparently" (p.10): the variety, depth, quality, annotation and accuracy of training datasets is a key factor affecting AI performance. With the aim of improving the auditability of AI systems, the European Union should create shared public datasets and environments for AI training and testing. The European Union shall then ensure equal access to those data, not least to create a level-playing field which could benefit European actors and prevent a small number of large private companies – particularly from third countries - from accumulating undue advantage by establishing 'data monopolies'. - Critical concerns raised by AI (p.11/13): it might be opportune to add two additional points related to 'algorithmic biases' and on 'deception through AI', an increasingly pressing issue given the potential use of AI to manipulate information through more and more sophisticated applications (i.e. speech	With reference to this chapter, Fondazione Leonardo Civiltà delle Macchine would like to share its observations and suggestions on the following specific points: - Architectures for Trustworthy AI (p.19): AI platforms should include an explicit and separated Software Module specifically responsible for "ethical subroutines". The modular design of systems might imply that no single person or group can fully grasp the way in which the system will interact or respond to a complex flow of new inputs. The same logic can be applied to "self-monitoring" modules that check system for behavioural consistency with the original goals. The inevitable by-product is the opportunity to separate the tasks of the data scientist, who will be responsible solely for AI logics, from those of the ethical scientist, who would be responsible for ensuring the above mentioned behavioural consistency. - Accountability (p.14): when dealing with the fundamental issue of accountability, it would be advisable to distinguish, wherever possible, accountabilities and liabilities between AI developers and AI users. To explain better the concept: a man is shot by another man with a gun, who is accountable? The man that shots the other man or the OEM of the gun? Where it is not possible to clearly allocate liability, it is necessary to develop a concept of shared responsibility. The impact of AI is the result of a multi-level system of interactions among designers, developers and users. AI is characterised by distributed agency, and thus by a form distributed responsibility. Existing ethical frameworks centred on the individual are ill equipped to deal with it. As argued by Taddeo and Floridi (2018), 'it is necessary to develop an ethical model that separates responsibility of an agent from their intentions to perform a given action or their ability to control its outcomes, and holds all agents of a distributed system, such as a company, responsible'. Furthermore, the issue of responsibility should not be evaluated within a vacuum, but always considered in light of a framework of 'social optimization'. The disruptive potential of AI systems operate on an aggregate level, improving the utility functions for which they are designed. For instance, the spread of autonomous driving vehicles will probably help decrease the total number of road accidents, an objective which we collectively think of as beneficial. In those cases in which the AI is not faulty and is designed according to commonly agreed standards, its liabilities should be purely assessed in terms of how well it optimizes its objectives. In a condition of force majeure, a perfectly designed and	With reference to this chapter, Fondazione Leonardo Civiltà delle Macchine would like to share its observations and suggestions on the following specific points: - Certification (additional, new point): If the High-Level Expert Group accepts our additional point pertaining certification in Chapter II. Realising Trustworthy AI, the assessment list of Chapter III could be expanded as such: • Has the company set up an internal committee for AI ethics? Or an organisational model that can "supervise" ethical aspects (e.g. "ethical assurance" similarly to "quality assurance") against an "ethical value chart – code of ethical conduct in algorithms' D&D" adopted by the company • Has the product undergone an ethics validation process by the internal committee for AI ethics? • If the product is set to perform critical tasks, does it comply with the rules and regulations of the SAFE-CA?	One of the chief missions of the recently established Fondazione Leonardo Civiltà delle Macchine* is to promote a new age of technological humanism, bringing civil society closer to the values of innovation and informing the general public on the potential of dual-use technologies in fields such as aerospace, defense, security and ICT. In this very moment, this specific mission is of primary importance for the Fondazione, considering that 2019 will also mark the 500th anniversary of the death of Leonardo da Vinci, one of the greatest creative minds of all time and forbearer of innovations that are still shaping several fields nowadays. For this reason, one of the Fondazione's main goals is to foster discussion and dialogue with civil society around the ethical and legal aspects connected with Research and Technology (R&T), particularly when it comes to groundbreaking developments like AI. This is why Fondazione Leonardo welcomes this report and consultation, which we hope will lay the ground for a EU human-centric approach to the development of AI applications capable of serving the common good and benefit the whole European society. * For more information about the activities of Fondazione Leonardo Civilita' delle Macchine, please refer to the following contacts: Segreteria Fondazione Leonardo – Civiltà delle Macchine Ph. +39 06/32473182 Web: https://www.fondazioneleonardo-cdm.com Mail. segreteria@fondazioneleonardo-cdm.com Add. Palazzo Grazioli, Via del Plebiscito 102, 3° piano

- An integrated approach: when it comes to the overall structure of the report, it is worth pointing out that It can be extremely difficult to translate ethical assumptions into procedural activities: nowadays designers finally start considering the ways in which they implicitly embed values in the technologies they produce, and a document like this one can definitely help engineers to become more aware of their work's ethical dimensions. But it is not so simple for engineers, given modern AI based systems' complexity, to predict how a system will act in a new situation. Engineers, companies, research centres, and design teams often work on individual hardware and software components that make up the final system. The modular design of systems can mean that no single person or group can fully grasp the way the system will interact or respond to a complex flow of new inputs. Maybe the best approach might be to let ethics only evaluate AI as the result of a pure engineering logic, rather than it (see "ethical subroutines" below).

- Data, Transparency and accountability: AI will increasingly leverage on algorithms and, to an even greater extent, on data. This situation raises two issues: (1) the transparency regarding the design of the algorithms and how these work, on one side, and (2) the quality and reliability of the data being leveraged, on the other. Ideally, and irrespectively of the critical nature of the tasks performed, trustworthy AI might not imply specific requirements for algorithm transparency when leveraged data are sorted from a "global common data" made up of publicly available open-source datasets. However, when such AI data are not sorted from this open-source "global common", absolute transparency on algorithms should be required, along with ethical and legal accountability for the OED (Original Engineering Algorithm Designers) and Service Operators (where applicable). Always in this context, a precise definition of critical and non-critical-task should be examined, so that appropriate regulatory frameworks could be devised for each of the various fields of AI application.

and imagery reproduction, etc.). This addition would also add coherence to the report, since both issues are mentioned in major reports on AI (Asilomar Principles, AI4 People) and in the recommendations in chapter II and III of this report.

In addition, when dealing with Covert AI Systems (5.2), the coverage should be extended to any deployment of AI, which interacts with any person within the EU. This includes all forms of 'active' man-machine interactions, such as chat-bots and androids, already mentioned in the draft report, but should also encompass 'passive interactions' where automated decision systems affect citizens and customers - e.g. automated price-setting and advertisement in the private sector, and welfare and social policy in the state domain. Particularly in the latter case, the lack of transparency in the application of automated decision-making to public policy can cause citizens' disenfranchisement and risk jeopardizing the weaker strata of society.

Regarding potential longer-term concerns (5.5), the possible future development of Recursively Self-Improving AGI and Artificial Moral Agents could usher in a deep re-thinking of existing ethical paradigms. AI systems designed for fast and recursive self-improvement should adhere to strict safety and control measures. Simple deontological ethics however, with its focus on fixed rules and moral duty, risk obstructing or slowing future growth of AI. On this point, the document should commit to a form of normative ethics anchored to consequentialism or Deweyan pragmatism. These two schools of thought focus, respectively, on the moral consequences of action and on the historical evolution of human society. They seem thus more apt to change and accommodate future technological development while keeping with the ultimate goal of human autonomy and freedom.

functional autonomous car might still cause harm to its passengers or to passer-by, but that would still be the least bad outcome. Such an approach shifts the focus of AI regulation from simple ex-post penalties to an ex-ante collective discussion on what the merit and aims of new technologies should be.

- Standardisation (p.21): given the novelty and new boundaries established by AI technologies and applications, standardisation should be extended to the following domains:

• Software engineering: especially for security, as well as to monitor and control emergent behaviours;
• Performance: to measure accuracy, reliability, robustness, accessibility, and scalability;
• Safety: to evaluate risk management and hazard analysis of systems, human computer interactions, control systems, and regulatory compliance;
• Interoperability: to define interchangeable components, data, and transaction models via standard and compatible interfaces;
• Cybersecurity: some cybersecurity risks are specific to AI systems. Those include, for instance, "adversarial machine learning", where AI systems can be compromised by "contaminating" training data, by modifying algorithms, or by making subtle changes to an object so to prevent it from being correctly identified; Assessing the future of cybersecurity is not an easy task: by its very definition, the introduction of new technologies can alter the material and social environment in which they operate, 'creating' the space for a new array of malicious attacks (i.e. a human shaped dummy hanging on the street does not represent a threat in a world dominated by traditional cars, but might represent a real DoS for autonomous vehicles)
• Traceability: to provide a record of events (their implementation, testing, and completion) along the entire AI system life-cycle.

- Stakeholder and social dialogue (p.22): in order to strengthen a constructive and healthy exchange between AI researchers, policy makers, stakeholders and civil society at large, the EU should create a permanent science-policy platform. At the same time, following the Asilomar Principles, the EU should nurture, through the European Research Council, a culture of cooperation and transparency among all the institutions of member states involved in AI research, fostering a constant dialogue and exchange of scientific information. Lastly, and most importantly, the EU should set up a comprehensive development plan to integrate AI into the educational space at all levels, from compulsory schooling to university and training courses for the workforce.

- Certification (additional, new point): as anticipated above in the "Data, Transparency and Accountability", another dimension to be countered is if AI covers critical or non-critical tasks. In such cases, different certification systems should apply, with the need to define what is critical and what is not under regulatory frameworks specific to each of the various fields of AI application.

					In the case of non-critical tasks, company-level certification should apply. Any company that deals with AI based software should provide its own ethical screening through a specifically appointed internal committee for AI ethics. The committee would be in charge of validating products before they are released to the market. This would allow a mechanism of corporate compliance with accepted ethical norms to help build user trust. Furthermore, it would make the development of AI algorithms less burdensome for software engineers, postponing the process of ethical assurance to the latter stages of product design. In case of any critical tasks performed either by general or narrow AI, a stronger public system of certification should apply. To facilitate this mission, the EU should create a Secure AI For Everyone Certification Authority (SAFE-CA), in charge of validating the compliance of critical AI assets with human rights and European regulatory norms. This would mitigate legal and business risk associated with business-to-business and business-to-regulator use of AI. Such Authority would also facilitate interoperability by providing a secure, enforceable, and regulatory-compliant way to verify the fulfilment of ethical and trustworthiness requirements.		
Jan	Wiesner	ARD (German Public Service Media)	Overall we support the statements on the 'Role of AI Ethics' (page 5), however we would like to emphasize the role of AI in regard to "freedom to live in a democratic society" (p. 5) as one of the central points. A democratic society certainly also need strong, pluralistic and free Media, which are increasingly impacted by AI. Therefore, as a general statement, the proposed guidelines should put more emphasis on the role AI play for our (democratic) societies and especially its possible impact on the freedom of speech in general and the media in specific.	On 3.2. Freedom of the individual: in the last sentence, the freedom of speech, i.e. to freely receive and impart information, should be included in the list of basic freedoms. On. 3.4. Third sentence: should read: "In an AI context, equality entails that the same rules should apply for everyone to access to and dissemination of information […]". On the Principle of Beneficence and the Principle of Non Maleficence (page 8f): We welcome the fact that the significance of AI for "the democratic process" and "freedom of speech" is recognized, as well as the statement that "AI systems should be developed and implemented in a way that protects societies form ideological polarization and algorithmic determinism." We have, however, a number of comments on these two parts of the guidelines: - Freedom of speech should be specified, entailing "freedom to receive and impart information"; - Freedom of expression and the danger of polarization are only mentioned under the Principle of Non Maleficence. We would, however, suggest to include these points – along with the 'protection of democratic processes' - also under the Principle of Beneficence. It should at least be considered what positive role AI can play in this field; - Regarding the ideological polarization, specific reference should be made to "fake news" and the problem of "information disorder" (see: https://www.ebu.ch/publications/perfect-storm). In this context the specific role of (public service) media, including the use of AI by the media, should be mentioned. On 5.2. Covert AI systems: We agree that a human should always know if she/he is interacting with a human being or a machine. In addition, we suggest to extent this requirement to the products or services resulting from the deployment of AI systems. This seems to be especially	On 1. Accountability (p. 14): The last sentence should read: "In case of discrimination or misinformation / Fake News, however, an explanation and apology and possibly a correction might be as least as important." On 6: Respect for (Enhancement of) Human Autonomy (p. 16): "extreme personalization", "manipulative 'nudging'" and "the role of recommender systems" are rightly highlighted as possible threats resulting from the use of AI systems. However, there should be a clear reference to News / Media in this context, as the mentioned problems are especially pertinent in this field. AI systems could be of major benefit for the Media, however they should be designed in a way that prevents AI to cause the identified problems. They should possibly even enhances plurality and diversity, supports freedom of speech and help to enable an open debate within society - all points that are of major importance for a well-functioning democratic society.	- no comment -	The proposed guidelines should put more emphasis on the role AI plays for our (democratic) societies and especially its impact on the freedom of speech in general and the media in specific.

					relevant in the field of journalism / news, regarding the results of roboter-journalism specifically in the field of news that (could) have an impact on opinion-forming.	
CCOO	Confederación Sindical de Comisiones Obreras	CCOO Spain				Artificial intelligence (AI) – or more accurately: automated decision-making - is already in use all over the EU, even if it is invisible. Often its workings are deliberately opaque in order to protect – open and hidden – corporate interests. AI is not just about technology or software programs, but societal choices are incorporated in this automated decision-making, for instance 'social scoring', credit lines. A debate about discrimination, equality, social justice, participation in relation to AI is needed. It should be clear that AI should not discriminate, it should strengthen equality, enhance social justice and participation. Such a comprehensive approach can't be limited to ethics. The debate needs contributions from sociology, philosophy, political science, economics and data experts. The focus of the discussion must be on the politically relevant questions – at national and at EU-level. CCOO welcomes the approach to connect AI with European values and principles. This is a first step in the right direction, but more steps are needed. New technology, and in particular Artificial Intelligence, must be shaped in way to avoid a threat to democracy and functioning markets. First and foremost, it has to be determined which of the challenges posed by AI can be addressed by enforceable rules and laws and which can be left to unenforceable ethic codes, guidelines, self-regulation or voluntary self-commitments. In modern democracies it must be a principle that its cornerstones, the principles of democracy, the rule of law and human rights, must from the outset by design be incorporated in AI.Citizens and workers, in particular workers' representatives in companies and public administration must be empowered to understand the new challenges ahead and be enabled to find appropriate answers. The GDPR was a first step in the right direction, but more regulation is clearly needed (for self-driving cars, face recognition, drones etc.) . The Commission should play a role to launch such a holistic debate involving a wide range of stakeholders and contribute to close the gap between Member States.The focus of CCOO lies in the world of work, in particular the future of work. AI needs to be embedded in decent work. AI is ambiguous and needs to be shaped, it can be used to cement power asymmetries or to dismantle them. It is in the interest of workers that information, consultation and board-level participation rights as well as collective bargaining are respected and fully applicable. A general information of stakeholders is clearly insufficient. The rights to information, consultation and board-level representation must cover the area of AI. A technological and social impact assessment is necessary as well as participative research to follow the design, application and

implementation of AI and its economic and social consequences. It is of utmost importance that enforceable regulation creates an appropriate framework for AI in Europe. AI cannot work in a lawless zone where chatbots are not identifiable, can contribute to hate speech, influence democratic elections and undermine democracy itself. In view of the upcoming European elections, but also in democratic discourse generally, it is important to know whether one’s counterpart is a human or a machine, which is not the case currently.The rules for AI are not yet in place and it is important to take the necessary steps.The respect for human rights, for workers' rights, for humans' moral and physical integrity, and the cohesiveness of our societies are fundamental goals, which cannot, and should not, be left to the free appreciation of businesses regarding their marketing or communication strategy. The reaction of the EU to the very real threats posed by AI to the achievement of these goals cannot restrict itself to indicative guidelines, with no external scrutiny and no sanction in case of non-compliance. CCOO thus demands strong, enforceable, regulation of AI, based on legislation. EU-wide legislation has the advantage of preventing downward regulatory competition among Member States.The legislation should prescribe procedural steps and institutions within organisations to ensure the trustworthiness of AI applications (under the model set by the GDPR), which can be verified by any layperson, and should limit to a maximum "ethics panels" or "boards", often self-serving, which do not provide sufficient predictability of their decisions and/or are vulnerable to conflicts of interest.This legislation should bear upon the following aspects, in addition to those already described in the document:* human workers must be able to take decisions different from the "recommendation" made by the AI system, and yet not be sanctioned for having done so when this decision proves to be wrong;* human workers must be able to test, experiment and innovate, even against the "recommendation" made by the AI system, and yet not be sanctioned for having done so when the test / experiment / innovation fails;* AI systems must be sufficiently reliable and their behaviour must be reproducible enough to ensure safety of material systems (specifically: of machines in a working environment), and particularly of "safety critical" systems where failure is known to cause deaths in large numbers (e.g. civil aviation, rail equipment, chemical plants, civil nuclear power);* AI systems must only be deployed in safety-critical applications after the level of explicability of the decisions, and the capacity to trace back an accident or incident to its cause, are sufficient for this cause to be treated, and for the safety of the application to improve over time; workers must be trained to deal with AI in particular to apply the emergency brake where necessary;* robots (aka "chatbots") must be identified and visibly marked in all on-line debates and discussions, so as not to be mistaken with genuine human opinions, or even be prohibited from taking part in some on-line discussions (e.g. on political, social or moral issues, in particular during election campaigns);* the added value created by AI must be distributed fairly in society and

							economy, specifically by making sure that the access to the data that teaches AI systems is broadly distributed among all economic players under Fair, Reasonable and Non-Discriminatory (FRAND) legal and economic conditions, and cannot be captured by digital monopolists. The requirement of "distributional fairness" must be added to the list of "Requirements of trustworthy AI" given in §II.1.
Yves	Poilane	Telecom ParisTech	Telecom ParisTech supports HLEG's proposed framework for Trustworthy AI "made in Europe", based on ethical purpose and technical robustness. 1. Encourage an approach open to international standards Telecom ParisTech believes that a "made in Europe" approach to ethical IA should incorporate wherever possible other international and regional approaches to ethical AI, including work done by the IEEE on Ethically Aligned Design (https://ethicsinaction.ieee.org). Given the global nature of AI innovation and research, we believe that a purely European approach risks weakening the impact and uptake of the HLEG's Trustworthy AI principles. International convergence should be sought wherever possible on AI ethics, bearing in mind that Europe has the potential to set the gold standard for AI governance globally (https://www.economist.com/business/2018/09/20/can-the-eu-become-another-ai-superpower). The sixth paragraph of Section A might be supplemented as follows: "This is the path that we believe Europe should follow to position itself as a home and leader to cutting-edge, secure and ethical technology. Europe's approach to Trustworthy AI should strive where possible to take into account other international and regional approaches to ethical AI, with the ambition of setting the gold standard for AI ethics globally." 2. Assessment criteria for Trustworthy AI Telecom ParisTech agrees that a tailored, context-specific, approach is needed to Trustworthy AI. Context-specificity depends in large part on the regulatory environment. The risks and regulatory framework associated with unethical AI in the banking sector are different from the risks and regulatory framework associated in unethical AI in autonomous weapons. The criteria used to measure desirable and undesirable	Chapter I: Respecting Fundamental Rights, Principles and Values – Ethical Purpose We suggest that this Section might usefully refer specifically to IEEE Ethically Aligned Design's discussion of: - the effect of AI on human well-being, a concept distinct from human rights and ethics (IEEE, Ethically Aligned Design, Version 2, pp. 24, 240-263); - the effect of AI on employment and sustainable development goals (Id., pp. 136-138).	Traceability & Auditability; Explainability Telecom ParisTech agrees with the HLEG's conclusions (p. 20) that traceability and auditability will require development of human machine interfaces that provide mechanisms for understanding the system's behaviour, and that regulatory bodies charged with overseeing AI systems will need to undergo a digital transformation and develop the necessary tools to verify and audit AI systems. This is why Telecom ParisTech believes that explainability, traceability and auditability need to be designed with the end objective in mind, which is to permit regulators to understand and measure parameters that are relevant for those regulator's particular missions. AI system developers, standards bodies, and regulators therefore need to work closely together to define appropriate toolboxes (KPIs, interfaces) for explainability, safety, and ethical AI performance in each sector. Regulation and standards We recommend that the HLEG Guidelines mention the distinction between design standards, which prescribe use of a particular technical solution to reach an outcome, and performance standards, which prescribe only the particular outcome, leaving the choice of technology to the developer. The distinction is important in the field of AI research, where different competing technological solutions may exist to achieve objectives such as explainability. Wherever possible, regulation and standards should refrain from prescribing use of a given technology, and instead define the outputs and measurement criteria used to assess them. This approach is generally better for innovation.	As mentioned above, Telecom ParisTech believes that assessment criteria for particular AI use cases should be built with the end objectives of regulators in mind. Trustworthy AI systems can be designed with these end objectives in mind, and include KPIs and measurement criteria designed to facilitate regulators' and auditors' tasks. For example, Trustworthy AI systems for autonomous driving might have separate explainability, traceability and auditability modules depending on which regulator, and which issues, are at stake. The explainability interface for privacy may be different from the explainability interface for collision avoidance, given the very different regulatory objectives and institutions involved. The HLEG assessment lists ask on several occasions how the relevant objective(s) should be measured and assured. For Telecom ParisTech, one of the most critical tasks for assuring Trustworthy AI will be to define standard KPIs and measurement criteria for desirable and undesirable ethical outputs, so that different systems can be evaluated and compared. However, human rights and ethical concepts such as "fairness" are hard to measure objectively. Recent research in AI (machine learning) show for example that it is difficult to implement fairness in algorithms because the concept of fairness accepts several definitions that can be incompatible between them. In addition, ethical concepts are based on societal norms that vary based on context and geography. Yet some form of measurement will necessary in order to assess and compare AI systems. In the field of law enforcement, one interesting experiment is the SURVEILLE project (https://surveille.eui.eu/), which consisted of developing a scoring method to compare different forms of police surveillance technology based on their impact on fundamental rights.	As noted above, Telecom ParisTech's main comments on the draft guidelines for Trustworthy AI are: - To emphasize international norms where possible, including the work of the IEEE on Ethically Aligned Design. The "made in Europe" approach for ethical AI will be stronger if it embraces international approaches where possible; - To focus more on the interface between regulators and AI explainability. Tools for explainability need to be built with each regulator's objectives in mind; - To emphasize the need for developing tools to measure and compare hard-to-measure ethical performance criteria such as AI "fairness".

			outcomes in AI systems, and explain AI decisions, will also be different. These criteria need to be designed in close collaboration with sector-specific regulators downstream, who will know what needs explaining, and what needs to be measured in AI systems in light of the public policy objectives those regulators are charged with defending. The last paragraph of section B(III) might be modified as follows: "…Given the application-specificity of AI, the assessment list will need to be tailored to specific applications, contexts, sectors and regulatory environments. The criteria used to measure and explain desirable and undesirable outcomes in AI systems will differ depending on the regulatory context. To develop explainable AI, engineers must know what needs explaining, to whom, and why. The criteria for measuring and explaining AI outcomes therefore need to be developed in close collaboration with sector-specific regulators, who will be in the best position to know what to measure, and what to explain in light of the public policy objectives those regulators are charged with defending. Tailoring the assessment list to regulatory environments is all the more essential that research shows that there is a negative relationship between the performance of AI (predictive accuracy) and its explicability. The most successful methods (e.g. deep learning) are often the least transparent, and the most transparent methods (e.g. decision trees) are sometimes less precise. We selected a number of use cases to provide an example…"					
Maciej	Groń	Polish Ministry of Digital Affairs	• Definition of AIWhen working on an AI definition, it is our view it is necessary to consider the place AI will have in the future. At the moment, AI systems operate in a separate IT environment, kind of a vacuum. In the not so distant future, that will change. AI systems will be comprehensively integrated into every network humans interact with - AI will for example become the essential component of IoT products. What is more, development of machine learning techniques will lead to the establishment of a very strong relationship between human and machine. Addressing this connection in the AI definition is tantamount to the establishment of a trustworthy AI. The definition of AI should acknowledge this fact and specify that AI is a technology which interacts with humans, both on the social and environmental level.Furthermore, with vast amount of data to process, AI systems will gradually move towards full automation. Therefore, in our opinion it is crucial for the AI definition to feature the notion of automated reasoning/decision-making.	• The Principal of Autonomy: "Preserve Human Agency"It is our view the principle of autonomy should refer to everlasting exclusivity of human control over machine, rather than focusing on human self-determination, defined as only knowledge of interaction with an AI system, plus right to opt out and withdrawal. The guidelines should clearly underline that when it comes to AI systems, the machine will always be the one to serve the human and cater to its needs. Never the other way around. By developing a human-centric approach to AI, we should always advocate for primacy of the human being in a context of technological change (as specified in the Ovideo Convention). This is very important in the long term. As the AI technology will evolve, human supremacy should be the inalienable foundation for further advancements made in this field. That is why, we would like to recommend including this line of argument either under the principal of autonomy or if possible by introducing an additional principle of human supremacy.• Principle of explicability: "Operate transparently"First of all, we fully support the initiative to introduce technological and business model transparency of AI systems. Knowledge on how this technology operates and more specifically, how a system has come to a decision is absolutely crucial to establish trust. We believe this principle should be strongly embedded into the EU socio-	• Data governance and human oversightHere we wish to commend the authors of the Guidelines for insisting on the need for high data quality, as proper functioning of AI systems depends heavily on it. As pointed out, datasets inevitably contain biases, errors, misrepresentations and other irregularities. Before such data is fed to the AI system it must undergo a thorough review and validation process. Same caution and control is advised when it comes to the process of data gathering and data diffusion. All in all, we are very happy data governance has been recognized as one of the requirements of trustworthy AI. It goes hand in hand with the necessity for human oversight of algorithms and data supply chain used in the design and development of the AI system (access to so-called "black-box"). As it has been depicted in the guidelines, as human beings we have the right to access information on the design and construction of a AI system, as well as on methods applied for the gathering and selection of components (data) used to make it work.	-	As a final point we would like to stipulate that we fully endorse the idea for Europe to become the leader of cutting-edge, secure and ethical AI technology. Trustworthy AI should be EU's trademark, what differentiates it from other systems. The goal should be for AI technology to be associated with systems that are secure, reliable and of high-quality, in other words which inspire trust.

			economic framework through education. Therefore, we would like to propose to expand the principle of explicability so that it includes the requirement to educate. In order to properly interact with an AI system every person must possess the necessary knowledge on its technological functionalities. This process of getting acquainted with the way AI systems work should thus begin at the earliest stage, preferably in school. Apart from equipping children with technological know-how, we should make sure they have the proper perception of the role AI system play in the society. Going back to the comment on human supremacy, every human being should be taught that he must always remain in control of the technology and that at no point should he relinquish it in favour of the machine. • Critical concerns raised by AI – Distortion of competition by means of data/information monopoly We would like to draw attention to a possible risk that may occur with the development of AI technology and which stems from unequal access to data. AI systems are fuelled by data and those entities which possess vast amounts of data enjoy a competitive advantage over their competitors. Such undertakings may use this fact to establish dominance in AI design and development and foreclose market entry. This in turn might distort competition on the market introduce a data/information monopoly. In order to provide a level-playing field for all European entities, efforts should be made to free access to data and ensure their flow across the EU and beyond in trusted ecosystem launched among like-minded countries.		
Anonymous	Anonymous	Anonymous	Take actions to prevent an „AI divide" on enterprise and employee level The disruptive development of AI technologies can turn into a „race to the bottom" between enterprises having the right means (digital maturity level, skill sets) to fully embrace the benefits and companies (depending on the sector, especially SME often lack a sophisticated digital architecture required for comprehensive deployment of AI technologies) with delayed adoption of AI and ultimately a lower share in the economic benefits of AI deployment. Another gap can unfold on employee level between workers disposing of AI relevant digital skills and those performing less specialized repetitive tasks. On the one hand, this might lead to an increased competition for the best talents and missing prospects for people lacking the required skills on the other hand. To prevent an „AI divide" on this level, strategies should be set up to support enterprises prepare and upskill their workforce in order to maintain employability in the age of artificial intelligence. Ethical codex for AI research AI research should be strongly guided by ethicals norms (as defined in this document) taking into account vulnerabilites and unpredictibel behaviour of AI. For any critical application (criteria for "critical applications" should be defined), one should be able to explain in a comprehensive way how the AI	Fostering of competive hardware and software „made in EU" High-performance and capable hardware and software is a precondition for trustworthy AI. This should be a priority in public funding schemes as from an international perspective, European countries are not well positioned in the market yet. Improving access to secure hardware infrastructre for SME, R&D and startups could be a lever to strengthen the European position in this regard.	Strengthen interdisciplinary partnerships on AI Interdisciplinary partnerships building upon competencies and information from innovations spaces, research think tanks, data-sharing platforms and business cooperations between large and small firms can help boost the European AI ecoystem while at the same time ensuring an multi-dimensional debate on the economic and societal impacts of AI. Open question: Measures or penalties for infringements of obligations relating to AI ethics? Effectiveness and power of control mechanims?	

				came to a specific result („Explainable and traceable AI"; also here, a common definition is required).		
Vittoradolfo	Tambone	Institute of Philosophy of Scientific and Technological Practice (FAST), University Campus Bio-Medico of Rome	The commentary has been performed by Marta Bertolaso (https://www.rd-alliance.org/users/marta-bertolaso), Laura Campanozzi, Nicola Di Stefano, Giampaolo Ghilardi and Vittoradolfo Tambone (FAST - University Campus Bio-Medico of Rome) in collaboration with Eugenio Guglielmelli and Loredana Zollo (Biomedical Robotics and Biomicrosystem - University Campus Bio-Medico of Rome)[p. I draft] "To ensure that we stay on the right track, a human-centric approach to AI is needed, forcing us to keep in mind that the development and use of AI should not be seen as a means in itself, but as having the goal to increase human well-being" (i).[our comment] In our mind AI doesn't need to be "human-centric" because it is an "Human Act that must be done in an ethical way": we need a philosophy of scientific and technological activity, not an ethics of technology because technology is not a free subject but only the product of human activity (free, intentional and for this reason good or bad). In other words, we think that the right Focus is not "How the AI must be" but "How the People must work in AI based systems".[our purpose] we think it would be useful to set the question as human centered, meaning that the real player of the incoming game will be researchers, engineers and all the people involved at some degree in AI based systems.[p. ii draft] "This document should therefore be a starting point for the discussion on "Trustworthy AI made in Europe". While Europe can only broadcast its ethical approach to AI when competitive at global level, an ethical approach to AI is key to enable responsible competitiveness, as it will generate user trust and facilitate broader uptake of AI. These Guidelines are not meant to stifle AI innovation in Europe, but instead aim to use ethics as inspiration to develop a unique brand of AI, one that aims at protecting and benefiting both individuals and the common good. This allows Europe to position itself as a leader in cutting-edge, secure and ethical AI. Only by ensuring trustworthiness will European citizens fully reap AI's benefits" (ii).[our comment] To realize a "Trustworthy AI made in Europe" document is a really fascinating goal, nevertheless it could be undermined by the adoption of a not European ethical model but rather of an USA one (Principialism of Beauchamp and Childress) , therefore we need to pay attention to what kind of ethics models we are going to choose.[our purpose] We are currently working on a shared European Axiological System as a collector of the main European Ethical Schools. If someone would like to join us in this job just email v.tambone@unicampus.it .[p. 1 draft] "The AI HLEG is convinced that AI holds the promise to increase human	[p. 5 draft] "These Guidelines are not meant to stifle AI innovation in Europe, but instead aim to use ethics as inspiration to develop a unique brand of AI, one that aims at protecting and benefiting both individuals and the common good".[p. 9 draft] Trust in AI includes: trust in the technology, through the way it is built and used by humans beings; trust in the rules, laws and norms that govern AI – it should be noted that no legal vacuum currently exists, as Europe already has regulation in place that applies to AI – or trust in the business and public governance models of AI services, products and manufacturers[our comment] Trust in a new Technology or in a new organization comes from understanding.[our purpose] This might imply, at the moment, at least just disentangling (1) trust in "internal" technological capabilities of the AI technologies, and (2) "external" trust in normative/institutional criteria for its application and use in the benefit of the societies and individuals. Together they contribute to the Human-Technology Interactions' Reliability within a "Human-Centric and Trustworthy AI" framework.[p. 8 draft] "The document states "AI systems should be designed and developed to improve individual and collective wellbeing" – we would keep coherence with the first part of the document, keep mentioning the common good". [our comment] Well-being is part of the common good, but is not sufficient to describe it.[p. 9 draft]. The document says: "by helping to increase citizen's mental autonomy." As already stated, we suggest using another kind of framework instead of that of the autonomy.[our comment] One of the challenges of living in a AI-based society is to keep one's own critical perspective on reality and ability to choose. [our purpose] Instead of the "mental autonomy", it can be rather helpful to state something like: "by helping to increase citizen's critical thinking."[p. 13 draft] "AI is human-centric".[our comment] AI features should not be limited only to ethics (fundamental rights, principles, and moral values), but also to aesthetics, as aesthetics represents a fundamental dimension of human life.[our purpose] To fully adhere to human life and to build a "reliable and trustworthy AI", the design should take into account also the aesthetics of AI systems, as it will deeply affect users' judgment. The document could also mention the concept of affordance, taking advantage of recent research on affordance based frameworks for human-machine interfaces and humanoid solutions.[p. 12-13 draft] "A balance must thus be considered between what should and what can be done with AI, and due care should be given to what should not be done	[Chapter II Draft] On the Principle of Explicability - 5.2 Covert AI systems[our comment] Here the document raises concerns about the integration of androids in human society[our purpose] In light of this section, we would recommend considering our proposal of calling this a "human-centric and trustworthy AI" instead of just a "trustworthy AI." This will help re-focus the attention on the primacy of humans over androids and humanoids.[p. 14 draft] "Requirements of Trustworthy AI (…)- Accountability- Data governance- Design for all- Governance of AI autonomy (human oversight)- Non-discrimination- Respect of (& enhancement of) human autonomy- Respect for privacy- Robustness- Safety- Transparency[our comment] As part of the robust technology challenge (and to some extent of the accountability challenge) are the criteria for ACCURACY and ADEQUACY that are missing in the document. ADEQUACY is structured by the question/goals we want to settle with our research programs or technology application ACCURACY: is determined in science by empirical evidence. The risk of 'opacity' related to a too wide and deep information for citizens should be considered (cfr. Refs)[our purpose] We would suggest to add RELIABILITY as an additional Trustworthy AI requirement to include ACCURACY and ADECUACY. Reliability deals with adequacy and accuracy because of its descriptive and explanatory dimensions that is, it refers both to the pragmatic but also ethical aspects of what we want to reach with what we learn and to the intrinsic consistency of the research program. We suggest to consider the impact that synthetic data –through AI technologies- will have on society, laws and communities. This will be particularly relevant for the insurances and the bio-medical fields (cfr. refs).[p. 14 daft] "II. Realising Trustworthy AI - This Chapter offers guidance on the implementation and realisation of Trustworthy AI. We set out what the main requirements are for AI to be Trustworthy, and the methods available in order to implement those requirements when developing, deploying and using AI, so as to enable full benefit from the opportunities created thereby".[our comment] All of this Chapter is interesting and seems to be the most important part of the Document. Unfortunately, it is not based on Evidence (if we understand well).[our purpose] We remark the purpose about the Survey of which we speak in our previous comment, because it could be useful to identify best evidence regarding: a) "additional technical or non-technical methods that can be considered in order to address the requirements of Trustworthy AI" [p. 22 draft]; b) "additional items to	[p. 27 draft] Fall back plan[our comment] This is a most crucial aspect that is, the off button. But in the end, according to the document, everything is resolved in the chance to restore a human control, so that it looks like making things too much easier for several reasons: when the problem is found, it could yet be too late.[our purpose] Hence, the robustness concept should be widening a little through for example the reliability concept (it will be suitable to manage future applications of AI to autonomous systems too) and to specify what protocols and control levels are to be implemented in the different cases.

wellbeing and the common good but to do this it needs to be human-centric and respectful of fundamental rights. In a context of rapid technological change, we believe it is essential that trust remains the cement of societies, communities, economies and sustainable development". [our comment] The accent posed in the Draft on the "human-centric" dimension of AI as essential is antecedent to the fact of being "trustworthy." Only a human-centric AI can be and needs to be trustworthy. [our purpose] We also endorse the importance of AI being "human-centric" instead of merely "human-centered" (cfr. Refs). To highlight the difference we can then speak of a "Human-Centric and Trustworthy AI" (ibidem). [p.3 draft] "The Guidelines are not an official document from the European Commission and are not legally binding. They are neither intended as a substitute to any form of policy-making or regulation" makes the Document weak".[our comment] It looks in contrast with what has been said in p. 19: " Ethics & Rule of law by design (X-by-design) - Methods to ensure values-by-design provide precise and explicit links between the abstract principles the system is required to adhere to and the specific implementation decisions, in ways that are accessible and justified by legal rules or societal norms. Central therein is the idea that compliance with law as well as with ethical values can be implemented, at least to a certain extent, into the design of the AI system itself. This also entails a responsibility for companies to identify from the very beginning the ethical impact that an AI system can have, and the ethical and legal rules that the system should comply with".[our purpose] We purpose therefore to present these statements as soft law.

with AI. Of course, our understanding of rules and principles evolves over time and may change in the future"[our comment] What kind of relationship can we establish with products of Artificial Intelligence?(1) In our opinion, this question is fundamental in order to give an ethical account of Artificial Intelligence and the possible answers are not clear in the present document.(2) Cfr. Refs. Various points of view underlie the connection between two polarities, called truster and trustee. However, the last one gathers the buried meaning considering the subjective substance.[our purpose] We propose an inquiry focused on the possible kinds of factual relationship in order to validate the ontological description made in the present document:(1) Developing further the ontology of trust and, then, the condition of possibilities (cfr. refs). (2) We cannot omit this duplicity if we want to analyze trust in A.I. Indeed, it is possible to interpret this expression in two different ways: the implementation in a cognitive system of rules that can be understood as trust, and the subjective belief that we can trust in products of A.I. To establish a deeper relationship between human and A.I., the way that we would like to bring ahead is to consider not only 'trust as reliance' or as a "consequence of decisions" (cfr. Refs) but as a "feeling" or "value", made not by a rational choice but derived also from an emotional statement.[p. 12-13 draft] "Realising Trustworthy AI" Ref. p. 1 draft: "Trustworthy AI - Artificial Intelligence helps improving our quality of life through personalised medicine or more efficient delivery of healthcare services. It can help achieving the sustainable development goals such as promoting gender balance, tackling climate change, and helping us make better use of natural resources. It helps optimising our transportation infrastructures and mobility as well as supporting our ability to monitor progress against indicators of sustainability and social coherence. AI is thus not an end in itself, but rather a means to increase individual and societal well-being. In Europe, we want to achieve such ends through Trustworthy AI. Trust is a prerequisite for people and societies to develop, deploy and use Artificial Intelligence. Without AI being demonstrably worthy of trust, subversive consequences may ensue and its uptake by citizens and consumers might be hindered, hence undermining the realisation of AI's vast economic and social benefits. To ensure those benefits, our vision is to use ethics to inspire trustworthy development, deployment and use of AI. The aim is to foster a climate most favourable to AI's beneficial innovation and uptake".[our comment] Trust is a key point; therefore it looks necessary to define it and describe it in a multidimensional way. In fact, in 2018 we worked on the hypothesis "Trust toward social robotics is related to inter-human trust". This idea has found empirical evidence on a pilot study "The robot I wish", which is ongoing under submission right now to the International Journal of Social Robotics. We are currently widening the sample of this research in collaboration with National Group of Bioengineering. We think it could be useful to perform the same kind of inquiry along the AI.[our purpose] According to what is reported on p. 22 of the draft: "We invite stakeholders partaking in the consultation of the Draft Guidelines to consider in order to ensure that the requirements for Trustworthy AI are implemented" [p. 24 draft]; c) "how the assessment list can be construed for and applied to the four use cases listed above, and what particular sensitivities these use cases bring forth that should be taken into consideration" [p. 28 draft].[p. 18-19 draft]: on critical concerns.[our comment] The data matter is really harsh for AI technologies (1) There are lots of recent approaches aiming at giving back to people their own data and selling them under payment only. But this is not about the right to be forgotten.(2) Actually, these data could never be absorbed by third parties but just sold time to time, even under payment if needed.[our purpose] (1) Regardless of whether this is not even mentioned, it seems to me that besides the right of opting out, the right to be forgotten is important, that is the right to request the cancellation of every own past data in order to make the opting out effective indeed. (2) One aspect not being predicted, but looking realistic to me, is standardising a consensus enunciation. That is i.e. being able to explicitly impede our own data being sold to a third party, and – even worst – triangulated with other sold by other data. (3) Looks like triangulation could most probably provide the most risky and subtle profiling, even if handled by governments and not only by private companies.[p. 22 draft] "Education and awareness to foster an ethical mind-set - Trustworthy AI requires informed participation of all stakeholders. This necessitates that education plays an important role, both to ensure that knowledge of the potential impact of AI is widespread, and to make people aware that they can participate in shaping the societal development. Education here refers to the people making the products (the designers and developers), the users (companies or individuals) and other impacted groups (those who may not purchase or use an AI system but for whom decisions are made by an AI system, society at large). A pre-requisite for educating the public is to ensure the proper skills and training of ethicists in this space."[our comment] We completely agree. In our mind, this education goal must be extended to primary and high school.[our purpose] We are thinking to realize an international School in Human-Centric AI & Social Robotics Applied Ethics and to study diversified training methods for young people. Whoever is interested in this project email g.ghilardi@unicampus.it .

				share their thoughts on additional technical or non-technical methods that can be considered in order to address the requirements of Trustworthy AI". We invite whoever wishes to participate to the new research to email l.campanozzi@unicampus.it . [p. 6 draft] "In short, fundamental rights provide the bedrock for the formulation of ethical principles" [our comment] Fundamental rights come from ethical principles, on the contrary they would be grounded in some extra ethical source. Particularly important is to avoid establishing fundamental rights on public morality because some time, the nazi experience on this is really clear, the majority is not synonymous of justice. [our purpose] We purpose to change the title at p. 5 "From Fundamental rights to Principles and Values" in "From Human Dignity to Principles and values". In this way, this Document could really understand Oviedo. In fact, when Oviedo said: "fundamental rights are the basic foundation to ensure the "primacy of the human being" it doesn't claim that Human Dignity comes from fundamental rights but only that the last ensures the first.		
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Luigia	Carlucci Aiello	Sapienza Università di Roma	This section is generally well conceived and understandable. I'd prefer to speak about AI systems rather than AI, but I do not want to enter into details on this.The document appropriately points out that the criticality in the use of an AI system heavily depends on the criticality of the application area of the system itself. The example of the recommender system for songs being much less sensitive than one for recommending medical treatment is very appropriate. This point could be brought a little further: the same AI system being used by two users at the same time may be beneficial to one and dangerous to the other one, depending on a bunch of circumstances. So, in general, being beneficial, or ethically acceptable, is not an absolute value of a system, but it has to be evaluated in a context.	%%%%%%%%Again, the presentation in this chapter is very accurate and good. Few notes and reflections: 1. The need for education is pointed out in Chapter III of the document. I believe it should be introduced much earlier as it is a fundamental issue in the development, and even more in the use, of AI systems. For instance, in Section 2, on page 5, the report says: "Informed consent requires that individuals are given enough information to make an educated decision as to whether or not they will develop, use, or invest in an AI system at experimental or commercial stages ...". My point is that educated decisions require information, but before that they require education, i.e. knowledge of the technology, its usage, effects, implications, and impacts on human lives. Education to prepare the new generations of developers of AI systems and, more important, education of the users. Conversely, without an adequate education, people won't be able to understand what is going on, won't be in the position of making an "educated decision'', and won't know how to "protect themselves" from dangerous or malicious uses of AI systems. This is already largely true, so actions are urgent. 2. In this chapter (actually in the entire	Through the entire document, in particular in Chapter II, it is evident that - despite the definition of AI provided in the accompanying document - AI is synonym of AI system, and that an AI system fundamentally is a Machine learning system, possibly a deep learning algorithm on a multilayer neural network. This is something very hard to agree with, and has the consequence of making the document incomplete in several ways.The importance of machine learning in the development of effective AI systems is out of discussion - its importance was already pointed out and accurately illustrated in some Turing papers in the forties of last century; the level of performance recently reached by AI systems due to important advances in the theoretical results and the technological advances and the criticality of the application domains of AI systems is also out of discussion. Nevertheless, this is only part of the picture, and this document should not ignore the rest, as there is more to machine learning, and there are many more research issues and technological challenges that contribute to the development of a trustworthy AI. The basic idea that informs this document is: AI gathers, collects, process, and update large amounts of data and then it makes decisions	Chapter III is a coherent and well developed list of all the items/issues presented in the previous chapters, so it shows the same strength and weaknesses.	A very well done document, the authors have to be commended for the huge effort made in organizing and presenting the very delicate and multifaceted problems connected with the development and use of AI systems. An incomplete document though, as it should include some other important components of AI systems and application areas that bring specific problem into the very articulated and varied scenario of the development of trustworthy AI devices.

				document) there is no distinction between AI technology and computer/digital/information technology. Even thought it is not at all easy to draw a border between them, it is important to point this out. First to tell people that here we are not speaking about something completely new, but technologies we have been living with since decades and that are evolving. Second, not to attribute to AI merits or demerits that should pertain to others. Subsection 3.5 is an example of an area where a lot can be done (or should have been done) without resorting to AI at all.3. The principles listed in Section 4 are easy to agree with, even though some of them are so "high level" to be of little use. The principle of "no harm" being an example.4. The principle of explicability brings me back to the observation that (a) education is very important, otherwise explicability becomes very difficult, and that (b) maybe it is better to speak about various types of explicability, e.g. one for the final user, and one for performing the acceptance tests of a system before its distribution/commercialization and use/maintenance. The level and kind of knowledge on the side of the humans and the level of detail and the kind of explanations provided by the AI system are necessarily different according to these circumstances.5. As for Subsection 5.5, I agree that the topic could be controversial. I suggest to leave the text in its present form, with an indication to "keep our eyes open" on the issues of General AI, consciousness, and full autonomy.	on the basis of some properties automatically extracted from these data. So, if we may trust the algorithm that governs the various aspects of this complex process, and we can trust the quality of the data, we are done: we have a trustworthy AI.My point is: all this is vey important, it is an essential component of an AI system, even though, it may be absolutely not easy to realize, but - more important - it is not the only component, as we need to know, among other things:1) the models of the world used by the AI system;2) the causal models underpinning the system, as Jude Pearl points out in his recent book, where he reminds us that correlation and causation are two tremendously different concepts; 3) the rules for inferring effects from causes;4) the possibility of speaking about 1) and 2) and 3) during the interactions between humans and AI systems, because this is the only way a human can understand WHY the AI system is behaving in a certain way and proposing/making a certain action.On page 19 of the document causal models (together with other very important issues) are mentioned en passant as something for future research and that possibly can inform the second deliverable. I'd prefer to see more in this document.Another limitation, that is possibly a consequence of the predominant machine-learning-on-big-dataspirit that informs the document, is in the sample application domains that are presented.I consider it important to address also other areas that are becoming very important and urgent, and raise their own specific issues concerning trustworthiness, namely robotic systems designed to act in particularly critical domains: personal assistants for the impaired, for the cognitive weak, for the elderly; robots used in the education of either children or adults. Here the issues to be addressed are sometimes very subtle, and I do not find any hint on them in the document.At the same time, I do not find in the document hints on very delicate and urgent topics such as computer-brain interaction, or prosthesis that enhance either the brain or the physical capabilities of individuals, or micro robots that navigate into the human body. More generally speaking: intelligent devices that interact with the human body with the aim of repairing /enhancing damaged or missing functionalities, or - why not - inventing new ones.		
Ozhan	SAGLIK	In my opinion, we should add the mechanism of certificating the products that use AI. By which means we can trust the products adopting AI is trustworthy (a certificate, a seal or statement approved by AI Alliance Consultation)	Archivists and record managers are keeping the information for decades. They are respecting human rights in their work and have produced theories about managing the information. Ethical principles stated in the Draft of AI Ethics can be adopted the theory of archival science and record-keeping theory. In archival science, it has defined that which records can access by whom and how which of them have critical value for the business or society, how the accountability and transparency of the process ensured and by which means the records can be preserved. We can use this knowledge on the principles of beneficence, non-maleficence, autonomy, justice and explicability.	Data science and archival science have some discrepancies but the theory of managing data is emerging now. We can not use record-keeping theory in every side of data science because data is fluid but records are stable. On the data governance, we can adopt new theories, technologies and practices but for the requirements of accountability, respect for privacy, robustness and transparency we can adopt record-keeping practices. Archivists are working on ensuring/protecting the reliability, accuracy, integrity and authenticity of the records. They are taking regulations, standards and society expectations so we must protect these features for AI products too so that we can benefit from the archival science and record-keeping.	We should decide by which means and by whom we can assess the trustworthy of AI, by seals that shows the AI of products' is trustworthy or certificate or another tool.	I would like to congratulate the members of an expert group. It seems they have worked elaboratively. I can offer to think my suggests about trustworthy seal or certificate.

Martijn	van Otterlo	Department of Artificial Intelligence and Cognitive Science, Tilburg University	It would be good to make more explicit what "European" values mean, possibly compared to American, Chinese, or Russian values, since AI is an international development, most of the ethical guidelines in this report are very broad, and at that level there is not much to go on to talk about "European" values in this part of the text. Page 4: why is there no arrow to "non-technical" methods? The rest of this section seems to make sense.	A similar question about the "European" values as for the introduction: can this be made more explicit? This chapter argues that it is best to build upon fundamental (EU) rights and derive principles and values from those rights. But, they again seem general, and the connection to the subsequent principles and values is still somewhat abstract. The example in 1.2 on going from "respect for human diginity" (right) to "autonomy" (principle) to "informed consent" seems plausible, but so would other translations into rights and principles too (e.g. relating to profiling, discrimination, manipulation, etc.). Again, the general line of reasoning seems fine, but often it stays at a very abstract level where many things would sound plausible. If the text wants to say that AI regulation (since that is one of the instruments) will be "embedded" into EU Law and based upon its principles, then this would sound plausible too, and maybe that is more to the point (provided the report would talk about "EU" values instead of "European" values). The diagram in Figure 2 is somewhat misleading. It seems that principles, values and right are at the same "level" whereas the accompanying text suggests other relations, relating to specialization, operationalizations and hierarchies. Section 1.2. but especially also 1.4. could do much more with the many other lists, guidelines, codes of ethics etc. that already exist for general technologies (including biotech, nanotech, nuclear technology, and so on), robotics, but also several other "AI ethical guidelines" such as the mentioned Asilomar principles (but see also Boddington's recent book). As far as I can see, this report brings (very) similar things to the table as other codes and guidelines, and it would be good to benefit from existing texts much more. "Explicability" as a term is interesting, but it is my impression that much of the AI (and bordering areas such as law) work on explainability already covers much the intended meaning by now (hence, not sure whether a new term is needed here, in this report). 5.3. Could add some more on pervasive systems monitoring and manipulating many kinds of interactions that were once non-digital. Plenty examples exist, like social communication, hiring&firing, shopping&paying, even reading. Having AI governing these interactions may change such interactions profoundly, beyond simple "influence". Also us relying on those systems has profound implications for "expert knowledge" (in humans) and what or who we believe (ranging from fake news to predictive models in companies and governments). For Section 5 there are (of course) many possible directions reported in the vast recent literature. For the end of Section 5, I would suggest adding a small piece on AGI, but then a different viewpoint. As Hugo DeGaris already predicted a long time ago, a clash could rise between people that would want to build general AI systems and those who would be	The section on trustworthy AI can be improved. The sense-plan-act cycle (and the specific way it is described here and connected to learning, stochasticity and so on) may not be the best example to explain this here. I do not see why the ethical goals and requirements should be integrated at the "sense" part (not: level as it is written). Relating to the previous point: what is really missing in this section on technical methods is "AI systems that can reason themselves about ethical behavior". Many researchers are investigating AI systems that can store, retrieve, learn, update, test, think about, and communicate, their ethical part of the decision making sub-system. So, one technical way in this section could/should be systems that are capable of making the right ethical decisions based on reasoning about the situations, and where norms and values can be taught, told, or built in. Another line of technical research that seems to be missing is the FAIR/FACT machine learning field, to make learning systems behave in a more ethical, fair, or otherwise trustworthy way. The main question in this section, which stay largely unanswered in this version, is some kind of guideline on how to go from highly general rights+principles+values, to very concrete principles,rules,values and decisions in actual systems. I see that with many high-level guidelines for AI: they all seem quite plausible, but the core question remains: how to "implement" them all the way down in actual AI machines? The explanation section is too narrowly defined. As it is written now, it seems that only AI implemented as a neural network should explain itself. Also the adversial examples are too much tuned towards neural networks. However, a huge amount of "explanation-oriented" AI work has been done over the last decades, and includes basically ANY system that does something meaningful, and for which one would like to know how it actually accomplished its task. This includes many tasks (including many NON-learning tasks) but equally so, many types of "representations" other than neural networks, including rules, many forms of logic, decision trees, propositional knowledge bases and so on. In addition, it would be good to note that explanations (and transparency) should be accompanied by two things: "what counts as an explanation" and "in what way does any explanation contribute to being more trustworthy"? There are typos scattered around the document, just to mention one here: footnote 28: "concpet" -> "concept".	The assessment list may be useful, but the question is: for whom? (designer, programmer, scientist, society) (and also: when, or in which stage, and with which consequence?)	This is an interesting report, and it gives a lot of useful information about how to build AI systems that have a chance of being trustworthy. As with many other texts on guidelines for AI (ethics) it stays at an abstract level, and the real work has only yet to begin: to operationalize the general guidelines, the "European values", and so on, and then actually build AI systems that --somehow-- contribute to these high level goals, and hopefully become "trustworthy". In addition, what could be added (more) to the report) is some remarks on "us", the humans, and how we perceive all these intelligent machines, and what "we" actually want. As the famous Moral Machine experiment has shown us, "ethics" is not something that is the same for all humans in the world, and presumably not for the whole of Europe, so if we want to build AI upon "European values", then maybe we should ALSO find out what "Europe" wants from AI?

			against it. Similar to (for example) biotechnology and nuclear technology, it can be expected that between-human clashes will rise for AI specifically too (for example, we have seen discussions in that directions for intelligent weapons, but maybe also now already with autonomous cars, where some people think we should not develop them in the first place). I think this would complete the story.				
Anonymous	Anonymous	Anonymous	Concretely to general commnets below:1. Declarative case „we" is frequently used in the document, which may be justifiable e.g. in constitutions of nations but not in guidelines that do not go via direct approval of all people in EU or even via their elected representatives. Expression commonly used in other guidelines (e.g. technical, medical) should be used in the document instead, e.g. „…we must ensure to follow the road that maximises the benefits…" (In Executive Summary) can be reworded „…the road that maximises …should be followed…"This questionable „we" can be found several times in the doc.2. Geographic issue. It is believed that that the team drafting the Guidelines knows geography but it is not clear why it uses Europe (most probably) for EU and also why randomly exchanges Europe and EU throughout the doc. Considerable part of Europe is not part of EU and does not follow EU legislation. It should be clear to which countries this document applies (probably EU or EU + EFTA?). When this is clear first, then the Guidelines can offer applicability of its provisions also to other countries in Europe, our neighbouring area. In Executive summary, there is sentence saying that the Guidelines aims to foster reflection and discussion on ethical framework „beyond Europe", but actually it overlooks number of nations in Europe as such that may be attracted to the document. 3. Figurative expression „north star" is at least 3 times in the doc in the sense that Trustworthy AI will (or is) „our" north star. Such expression is not very suitable for using in guidelines and might confuse its users from the AI industry. Moreover in other languages of EU countries the expression „to be North Star" (in other literature written with capitals) is not common and may even look funny. What do you actually want to say? e.g.: Is trustworthy AI an ultimate goal of EU countries?	1. page 10 Key guidance List of examples of vulnerable persons should be preferably kept consistent throughout the doc. (here elderly persons are missing).	1. page 15 Design for all. It should be clear that this requirement should be applied with respect to actual purpose or application of the AI system in question. The guidelines should not discourage suppliers and users of AI designed for narrow purpose and selected users only. This requirement should clarify this by a suitable wording e.g. First sentence of clause 3. „Systems intended for general public should be designed in a way that allows..." Plus additional sentence for the other (numerous) AI systems „This requirements can be adequately applied also to other systems intended for selected user groups only, depending on the purpose of such AI systems. " 2. Respect to privacy in part II. : The following requirement is not clearly represented in the list: All processes shall eliminate the risk of both AI providers and users getting access to the undisclosed test data and anyone getting access to the AI provider's intellectual property and technology. 3. Testing & Validating In addition the validation it would be useful to stress that AI systems intended for critical applications that influence life or health should comply with independent and transparent benchmarking system, when available. 4. Consider some provisions related to Performance, e.g. as an 11th requirement in part II. Even though performance may be subject of "other document". May a general requirement be formulated (?) E.g., assessing the performance of the AI implementation before they enter the market. This includes, definition of API interfaces, protocols and data formats for testing the AI systems.	1. Though it is mentioned somewhere in part II of the doc that regulation converting liability is subject of „second deliverable" it should be stressed in the front part of the part III. that part III. does not cover such issues. E.g. a sentence in the introductory section of part III should inform the reader that regulatory requirements are not covered in the list of requirements listed below. 2. page 24, similar clarification as in the comment 1. above (part II) should also be present in the list of requirements, possibly as a general statement. E.g. "The requirements should be applied with discernment to intended purpose of the AI system and expected users." 3. Compliance. Clearer expression related to demonstration of intended performance in comparison with other state-of-the –art solutions: Pls consider adding: Validation of AI systems that are planned to be used in applications related to health and safety shall be an important part of the system development process.	General comments: Considering the Guidelines are expected to be used by people developing and using AI, the Draft uses colloquial expressions that rather unusual if not improper in such kind of document. Also, it is felt that important issues of AI systems are mixed with less important ones, which may discourage the reader to overlook the important requirements.

Anonymous	Anonymous	Anonymous	Concretely to general commnets below: 1. Declarative case „we" is frequently used in the document, which may be justifiable e.g. in constitutions of nations but not in guidelines that do not go via direct approval of all people in EU or even via their elected representatives. Expression commonly used in other guidelines (e.g. technical, medical) should be used in the document instead, e.g. „…we must ensure to follow the road that maximises the benefits…" (In Executive Summary) can be reworded „…the road that maximises …should be followed…" This questionable „we" can be found several times in the doc. 2. Geographic issue. It is believed that that the team drafting the Guidelines knows geography but it is not clear why it uses Europe (most probably) for EU and also why randomly exchanges Europe and EU throughout the doc. Considerable part of Europe is not part of EU and does not follow EU legislation. It should be clear to which countries this document applies (probably EU or EU + EFTA?). When this is clear first, then the Guidelines can offer applicability of its provisions also to other countries in Europe, our neighbouring area. In Executive summary, there is sentence saying that the Guidelines aims to foster reflection and discussion on ethical framework „beyond Europe", but actually it overlooks number of nations in Europe as such that may be attracted to the document. 3. Figurative expression „north star" is at least 3 times in the doc in the sense that Trustworthy AI will (or is) „our" north star. Such expression is not very suitable for using in guidelines and might confuse its users from the AI industry. Moreover in other languages of EU countries the expression „to be North Star" (in other literature written with capitals) is not common and may even look funny. What do you actually want to say? e.g.: Is trustworthy AI an ultimate goal of EU countries?	1. page 10 Key guidance List of examples of vulnerable persons should be preferably kept consistent throughout the doc. (here elderly persons are missing).	1. page 15 Design for all. It should be clear that this requirement should be applied with respect to actual purpose or application of the AI system in question. The guidelines should not discourage suppliers and users of AI designed for narrow purpose and selected users only. This requirement should clarify this by a suitable wording e.g. First sentence of clause 3. „Systems intended for general public should be designed in a way that allows..." Plus additional sentence for the other (numerous) AI systems „This requirements can be adequately applied also to other systems intended for selected user groups only, depending on the purpose of such AI systems. " 2. Respect to privacy in part II. : The following requirement is not clearly represented in the list: All processes shall eliminate the risk of both AI providers and users getting access to the undisclosed test data and anyone getting access to the AI provider's intellectual property and technology. 3. Testing & Validating In addition the validation it would be useful to stress that AI systems intended for critical applications that influence life or health should comply with independent and transparent benchmarking system, when available. 4. Consider some provisions related to Performance, e.g. as an 11th requirement in part II. Even though performance may be subject of "other document". May a general requirement be formulated (?) E.g., assessing the performance of the AI implementation before they enter the market. This includes, definition of API interfaces, protocols and data formats for testing the AI systems.	1. Though it is mentioned somewhere in part II of the doc that regulation converting liability is subject of „second deliverable" it should be stressed in the front part of the part III. that part III. does not cover such issues. E.g. a sentence in the introductory section of part III should inform the reader that regulatory requirements are not covered in the list of requirements listed below. 2. page 24, similar clarification as in the comment 1. above (part II) should also be present in the list of requirements, possibly as a general statement. E.g. "The requirements should be applied with discernment to intended purpose of the AI system and expected users." 3. Compliance. Clearer expression related to demonstration of intended performance in comparison with other state-of-the –art solutions: Pls consider adding: Validation of AI systems that are planned to be used in applications related to health and safety shall be an important part of the system development process.	General comments: Considering the Guidelines are expected to be used by people developing and using AI, the Draft uses colloquial expressions that rather unusual if not improper in such kind of document. Also, it is felt that important issues of AI systems are mixed with less important ones, which may discourage the reader to overlook the important requirements.
European and international policies area	CGIL	CGIL - Confederazione Generale Italiana del Lavoro	In the context of a text that is rightly concerned with addressing risks of the artificial intelligence on the ethical sphere, we underline the importance of this passage. The development of Artificial Intelligence will mark the competitive position of Europe in the world, in a scenario in which the US are a leader, thanks to colossal investments in public research since the 70s, that have helped to create very few powerful multinationals and where China has announced its intention to become a leader by 2030 - and the very fast rates of development and the huge base of users (that means availability of data to feed the AI) can make this happen soon. Europe can not be left behind on this match, but rather it must propose a model of innovation - therefore of development of the AI - focused on the development of the technologies useful to respond to some of the biggest challenges of our times, from the environmental to the demographic and social one. The legal instrument for implementation should be clarified, since – in the already mentioned context – there is the risk that European countries are reduced to organize	Given the fact that the European Union is constitutionally committed to the respect of the fundamental human rights and that this remains an indisputable point of reference, in a historical moment of great social and political conflict, in which some of the dominant ideas are being questioned, the ethical debate must also be positioned in a critical way. The idea of "fairness" - fairness, justice, but also that of the common good - is a conflictual and dynamic concept in itself. There is the risk that in these systems it is established a non-negotiated balance of the idea of "fairness", which would become a static and objective data. We need to counteract a static and one-dimensional idea of fairness and to recognize that the values on which the algorithms are built, are not neutral. It is possible to do this with two precautions: 1) ensuring the greatest possible diversity while designing these systems, in terms of demographic variables but also of political and social representation (this is actually recognized in the text, where we talk about freedom from bias, stigmatization and discrimination); 2) recognizing the role of confrontation between the parties involved (in the case of	It is important to decline these points also with respect to the application of the AI to the organization of work. First of all we must recognize the importance of the data supplied to the artificial intelligence algorithm: despite being able to learn, the machine will always do it on the basis of the data supplied to it, therefore the collective negotiation of the datasets (in the case of work, but the same goes for participatory governance in other fields) becomes central and strategic. Always talking about labour, not only the individual worker must be able to decide autonomously about the solutions proposed by the machine (in order to avoid the creation of a new kind of alienation), but it is necessary to establish collective occasions for the ongoing verification of the effects of artificial intelligence, since it is not possible to know the outcomes of the learning process at the moment of its design. This is why, also in this case, the point on the Robustness is central, in particular the one on Reproducibility, which means reproducibility of the decision-making process: in fact, it is asked that the reasons and the process that led artificial intelligence	In addition to the four use cases, the Expert Group will investigate - (1) Healthcare Diagnose and Treatment, (2) Autonomous Driving/Moving, (3) Insurance Premiums and (4) Profiling and law enforcement - this Assessment list can also be adapted to the work field, both for bargaining and other participatory methods, in particular of organizational character. We are available for discussion, if the Expert Group is interested.	It is acceptable that the central role of the code is human dignity guaranteed also in terms of physical, psychological and financial security. And with the goal of the implementation of man's autonomy. It would be appropriate to indicate among the elements of guarantee to ensure equal access to all human beings to the benefits of AI and forbid its use for war purposes. The possibility that AI is reliable and secure needs an algorithmic construction that provides a continuous, ongoing human feedback of the self-implementation from machine learning, being some of which unpredictable at the origin. It is uncertain, in fact, the possibility of forecasting what is the "ethically" cheaper choice for AI in front of of different options choice. Generally speaking, the definition of the objectives to be pursued never protects from unexpected events but it is not impossible, even if in general what happens depends on the data that are input to "feed" the intelligence of the machine: it is important though to ensure the principle of Reproducibility as already stressed above. The consequences on the organization of work will be paradigmatic in this sense and the theme is not referred to in the code. In

philosophical conferences and debates, rather than engage on a advanced path towards development, in which the involvement of the social partners in innovation is paramount. We think a Directive or rather a Regulation is needed, as in the case of the GDPR, which recognizes the role of collective bargaining (not just information and consultation rights) in advance and in ongoing progress (in the development of AI systems, which are able to learn and therefore are constantly transformed) involving the workers.

work, bargaining between the social partners). This is needed to achieve a signification of the concept of fairness that takes considers also the point of view of the weaker party (in the case of a power relationship), of the minorities or otherwise of those who do not hold a cultural hegemony. The latter point should be more explicitly included already in the part on the Ethical Principles.
Point 5.4 Lethal Autonomous Weapon Systems (LAWS) We emphasize the importance of this point and the need for an international Treaty to ban the use of AI for war purposes.

to make a specific decision can be checked, through the reproduction of the same decision. In a context in which the bargaining part (or the stakeholders in charge of the choices of an algorithm) has not (yet) high skills in programming, it is necessary that once the values and inputs of the algorithm (also "objective" criteria, such as shift patterns) have been established, workers can map the results and see if the algorithm actually respects what has been agreed.
This is true for all the cases in which the inputs are subject to bargaining, so as to allow the ex post verification. Otherwise the risk is that the intelligence of the machine becomes an alibi to justify unforeseen behaviors, and make previous behaviors impossible to be known.

About Transparency, Transparency means training for everyone on the fundamental mechanisms that regulate digital and artificial intelligence, starting from the primary school but also for the adult population.
As indicated in the ETUI Foresight brief n. 05 (Aída Ponce Del Castillo, Artificial intelligence: a game changer for the world of work, June 2018), "This involves learning to work alongside AI and anticipating and visualizing how AI can and will transform their career and role in a company. This 'AI literacy' requires computer literacy, understanding, processing and manipulating data (and understanding its limitations), identifying and solving AI-related problems, logical and computational thinking, and generally acquiring the ability to live and evolve into a new (AI) world".

About Testing& Validating, we underline the relevance of this passage.

About Non Technical methods, We think it would be appropriate to add a point on collective bargaining, as one of the main tools of non-technical regulation: in fact, the scope of Application of the AI to the organization of work is one of the most relevant for the public and collective interest and participatory governance appears to be one of the most effective antidotes also to ethical risks. As indicated in the ETUI Foresight brief n. 05 (Aída Ponce Del Castillo, Artificial Intelligence: a game changer for the world of work, June 2018), “Those of the ‘open code movement’ believe that the code is the most ‘transparent’ part of the algorithm and should be accessible and open. Some data scientists, meanwhile, argue that the code may be useful but that what also matters is the data Idea Diffusa - Report della Discussione - Draft AI Ethics Guidelines For Trustworthy AI fed to the algorithm, the way it is selected and the form it takes”. it is selected and the form it takes ". We think that both paths must be followed, choosing the relevance of the methods also on the basis of the specific application context. This way the workers can ask to include other independent variables or attention thresholds or automatic suspension. In fact, in an enterprise an AI system could suggest that the business decisions have the maximization of the capital gain as an independent variable of capital: if the system proves to be quick, effective and able to learn from mistakes, the capitals will target

fact, as far as the worker will also be allowed, according to the the principle of transparency and understanding, to the knowledge of the functioning of the system, there is no specific provision for possible intervention by the acknowledged stakeholders (in particular the trade unions) to determine the amendment in case of unequal application outcomes and there is a lack even of the idea of an automatic suspension in case of certain critical issues. For this reason, we propose to add in the “Non-technical methods to achieve Trustworthy AI” a paragraph on collective bargaining, organizational participation and in general, on the involvement of the principal stakeholders involved.

					the companies that will adopt it and a few managers will take the responsibility of refusing to accept the suggestions. This is why it is important to include in the algorithms the principles of law enforcement and collective bargaining references, as well as contracting data and codes, according to what has already been expressed in the previous paragraphs.		
Anonymous	Anonymous	Anonymous	In general the Guidelines is acceptable.	Basically it covers everything well.	Basically it covers everything well.	Both Governance and local "factory" control is very important and unmissible.	It is important to detail more the definitions of the National decision-making process in the application.
Maud	Sacquet	Computer and Communications Industry Association (CCIA Europe)	We fully support the scope of the Guidelines, as drafted in this section. It is important to acknowledge that "different situations raise different challenges" and that tailored approaches are needed "given AI's context-specificity". We also strongly support the reminder that "no legal vacuum currently exists, as Europe already has regulation in place that applies to AI" and that public institutions (e.g. governments) are included in the target audience of the Guidelines as developers and users. On Trustworthy AI: We would suggest adding a consideration around the fact that building trust also means demystifying some of the unfounded concerns around the technology and educating the public on what AI is and how it can be used. On the Purpose and Target Audience of the Guidelines:The introduction of an endorsement mechanism seems unnecessary, given the voluntary nature of the Guidelines. It raises the question of whether there is a risk that the Guidelines may subsequently be referenced, e.g. in procurement procedures. In addition, the voluntary nature of the Guidelines is not properly reflected in the entire document. In particular, Chapter I states that "the section can be coined as governing the ethical purpose" and "identifies the requirements for trustworthy AI" – rather, it should speak of "providing guidance". The document puts forward a set of guiding principles and suggested practices – not mandatory requirements. We also should not imply that following these guidelines is the solution to deliver trustworthy AI, as the guidelines themselves mention that AI cannot be a box ticking exercise. We suggest reviewing the document to strike the right tone.On the Glossary:The definition of "bias" seems to overly focus on the human element. We would suggest the following, more nuanced, approach: "Bias is a prejudice for or against something or somebody, that may result in unfair decisions. It is known that humans are biased in their decision making and that unfair bias permeates our societies. Since AI systems are designed by humans and rely on data, it is possible that their results are, even in an unintended way. Many current AI systems are based on machine learning	On point 2 "From Fundamental rights to Principles and Values"The concept of "informed consent" is strongly linked to the EU General Data Protection Regulation (GDPR) and to consent to data processing. We would advise against using this expression in this context in the first place, to ensure that the guidelines are as clear as possible. If that's not the high level expert group's view, we suggest highlighting the fact that "informed consent", in the context of AI, is a broader point. This could be done as follows:"In turn, informed consent is a value needed to operationalise the principle of autonomy in practice – for example in case of a patient deciding to undergo a clinical trial. In the context of AI, informed consent would requires that individuals are given enough information to make an educated decision as to whether or not they will develop, use or invest in an AI system at experimental or commercial stages […]".On point 3 "Fundamental Rights of Human Beings"On 3.1. (respect for human dignity), we suggest to remove "rather than merely as data subjects" as this insinuates a negative attitude in the technology sector towards customers or citizens. Being a data subject has no negative connotation of its own – the issue arises only when data subjects' rights are not respected. On 3.3 (respect for democracy, justice and the rule of law), we believe the guidelines are not clear in their use of 'rights'. The document rightly refers to clearly defined rights as the guiding star to develop trustworthy AI. However, it also mentions different ones that are not codified. In addition, we suggest to replace "must not interfere" with "should serve to further democratic processes" for the same reason as in 3.1; and "a right to" with "an opportunity for". See the paragraph below with the proposed changes:"[…] AI systems should serve to further democratic processes or undermine the plurality of values and life choices central to a democratic society. AI systems must also embed a commitment to abide by mandatory laws and regulation, and provide for due process by design, meaning an opportunity for a human-centric appeal, review and/or scrutiny of decisions made by AI systems."	On point 1 "Accountability", it is unfortunate that the text does not acknowledge the accountability processes likely implemented within the organisations developing or deploying AI systems. We also suggest to add "accountability might include the ability to contest the output and provide feedback on why a certain result is right/wrong" to address the question of the quality of data sets. See the paragraph below with the proposed changes:"[…] In a case of discrimination, however, an explanation and apology might be at least as important. Accountability might include the ability to contest the output and provide feedback on why a certain result is right/wrong." On point 2 "Data Governance", it is unfortunate that the Guidelines do not mention established best practices, such as the traceability of data sources and data transformations and documentation on the quality and nature of data. Some claims should also be clarified or at the very least discussed, such as "biases can be pruned away before engaging in training" (contradicting a later statement underlining that "data always carries some kind of bias") and "it is advisable to always keep record of the data that is fed to the AI systems" (as this might not always be compatible with the GDPR). It is also worth noting that biases can also be corrected post-training (and not only in the training data or during model training). See the paragraph below with this proposed change:"The datasets gathered inevitably contain biases. One way to address them is to prune biases away before engaging in training. Biases may also be corrected in the training process itself by requiring a symmetric behaviour over known issues in the training set. Or they may be addressed post-training by adjusting how trained AIs are used, for instance by varying the thresholds used to convert model scores into decisions."On point 6 "Respect for (& Enhancement of) Human Autonomy", it is important to clarify that the concept of autonomy is a much bigger concept than B2C personalisation online and that a personalized shopping recommendation cannot be equated with practices that would harm humans' right to self-determination.	As a first general comment, the added value of this chapter seems questionable. The selection process of case studies appears somewhat unclear and arbitrary. At the very least, the suggested changes below should be taken into account. On point 1 "Accountability":As this first point is about "Accountability", we suggest to delete references to "responsible AI training" and "ethical oath", as this has little to do with accountability or redress. We suggest to move the point on diversity and inclusiveness to "Design for all" for the same reason. See the final list of bullet points below with the proposed changes:• What is the framework for redress if things go wrong?• Is everything in place to ensure procedures are followed and demands are met?• Can third parties or employees report potential vulnerabilities, risks or biases, and what processes are in place to handle these issues and reports? Do they have a single contact point to turn to?• Is an (external) auditing of the AI system needed and foreseen?• Has an Ethical AI review board been established? A mechanism to discuss grey areas? An internal or external panel of experts?On point 2 "Data governance":We suggest to clarify the first and third bullet points. See the final list of bullet points below with the proposed changes:• What process and procedures were followed to ensure proper data governance?• Is an oversight mechanism put in place? Who is ultimately responsible?• What data governance regulation and legislation are applicable to the collection of data and the particular use of the AI system?On point 3 "Design for all":As mentioned above, we suggest to add the bullet point on diversity and inclusiveness from the paragraph on accountability here. We also suggest to add a bullet point on the purpose of the system and its target users, as well as to remove the bullet point on "equitable in use" because this is a very high-level question, hard to answer on a practical level. See the final list of bullet points below with the proposed changes:• Does the system accommodate a wide range of individual preferences and abilities?• Is the system usable by those with special needs or	We support the constructive approach of the Guidelines, as well as their high-level focus. We support as well the fact that AI is clearly identified as a net positive for society. However, we are concerned with the negative tone regarding AI that appears often in the Guidelines, with a guidance often focusing on "do not" instead of "do". We are also concerned about the references to "AI made in Europe" and ethics as a competitive advantage, as the global nature of AI technology cannot be ignored. The best societal outcome is to boost ethical development and use at a global level. As explained in our comments above, we are also concerned about the reference to informed consent (strongly linked to GDPR, when in this case "informed consent" would have a broader meaning) and giving people a blanket right to refuse being subject to AI technology. In many cases, such a right could not be implemented, would go against the benefit of the user or against the rights of others. It could also impede the functioning of public institutions.

data-driven techniques. Therefore bias can manifest itself in the collection and selection of training data. If the training data is not inclusive and balanced enough, the system could learn to make unfair decisions. At the same time, AI can help humans to identify their biases, and assist them in making less biased decisions.“Of note: the definition of “trustworthy AI” seems clearer in the Glossary than the one used in the Executive Summary. In particular, it underlines that fundamental rights and regulations should be complied with during the development, deployment and use of AI. It is not the AI system itself that respects these.

On 3.5 (citizens rights), it is important to note that most government and commercial service provisions will in the future entail some degree of automatic processing of data. It is unclear how a blanket opt-out option would work in practice. Would a citizen, for example, have the right to systematically require the entire manual processing of his/her tax declarations? Therefore, we suggest to delete “and systematically be offered to express opt out. Citizens should never be subject to systematic scoring by government”. We suggest also to replace “hold potential to improve” by “are already improving” for the same reason as in 3.1. See the paragraph below with the proposed changes:“[…] AI systems are already improving the scale and efficiency of government in the provision of public goods and services to society. At the same time, citizens should enjoy a right to be informed of any automated treatment of their data by government bodies. Citizens should enjoy a right to vote and to be elected in democratic assemblies and institutions.”On point 4 “Ethical Principles in the Context of AI and Correlating Values”As the five ethical principles mentioned might be conflicting at times, we suggest to replace “must be observed” with “developers and users should strive to observe”. We also think it’s important to mention that one principle might risk coming at the expense of another. For example, strict privacy requirements might come in the way of more detailed – and fairer – datasets. See the paragraph below with the proposed changes:“Building on the above work, this section lists five principles and correlated values that developers and users should strive to observe to ensure that AI is developed in a human-centric manner. […] It should also be noted that, in particular situations, tensions may arise between the principles when considered from the point of view of an individual compared with the point of view of society, and vice versa. Also, it should be noted that one principle might risk coming at the expense of another. There is no set way to deal with such trade-offs.”On “the principle of non-maleficence: do no harm”, we suggest to replace “protects societies from” with “limits the risk of” and to delete “AI specific harms may stem from the treatment of data on individuals (i.e. how it is collected, stored, used, etc.) […] discrimination, manipulation or negative profiling.” As AI specific harms may come from many different sources, it is important to not always focus only on data collection and/or profiling if these Guidelines are to be applicable to all economic sectors. See the paragraph below with the proposed changes:“[…] At the very least, AI systems should not be designed in a way that enhances existing harms or creates new harms for individuals. Harms can be physical, psychological, financial or social. Of equal importance, AI systems should be developed and implemented in a way that limits the risk of ideological polarization and algorithmic determinism.”On “the principle of autonomy: preserve human agency”, the sentence on “direct or indirect AI decision making” and the right to opt-out and of withdrawal lacks clarity. For instance, how would work a right of withdrawal with an indirect interaction with AI systems? We suggest, at the very least, that the right of withdrawal applies “according to the use

The meaning of “extreme” and “nudging” is also unclear. Therefore, we suggest to add “the concept of autonomy has been discussed at length in the previous sections” at the end of the first paragraph and to delete the first two sentences of the second paragraph. We suggest also to replace “provide explicit support to the user to promote her/his own preferences, and set the limits for system intervention” with “respect their right to human determination”. See the paragraph below with the proposed changes:“AI systems should be […] autonomy of individual users and communities. The concept of autonomy has been discussed at length in the previous sections. Systems that are tasked to help the user, must respect their right to human determination, ensuring that the overall wellbeing of the user as explicitly defined by the user her/himself is central to system functionality. “ On point 8 “Robustness”, it is unclear why the question of transparency on the level of confidence or uncertainty with which predictions are made, included in an earlier version of the Guidelines, has been removed. On point 10 “Transparency”, we strongly recommend to introduce some nuance. A transparency requirement for “all models that use human data or affect human beings or can have other morally significant impact” is a very strong statement and lacks clear definitions. It is also unclear how such a statement would be applied to self-learning systems. Existing rules on the use of data should be the framework used for AI. We suggest to amend the second part of paragraph as follows:“[…] Explainability – as a form of transparency – entails the capability to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environments, as well as the provenance and dynamics of the data that is used and created by the system. Providing meaningful information about choices and decisions concerning data sources, development processes, and stakeholders should be required for uses that can have significant impact.” On Traceability & Auditability (p19), it is unfortunate that “transparent” and “understandable” are not better defined, as this is key for a practical use of the Guidelines. However, we fully support the acknowledgment that “the development of human-machine interfaces that provide mechanisms for understanding the system’s behaviour can assist in this regard”. On Explanation (XAI research), it’s important to note that the difficulty to provide clear reasons for the interpretations and decisions of the system is a known issue with learning systems based on neural nets but also when they are based on other complicated models. On Standardization (p21), it is doubtful that a unified horizontal standard could be meaningful and applied to APIs and interfaces, in addition to AI systems. It is also unclear what is meant by “standardization”. We would at least suggest to include “however, the nature of AI makes it difficult to imagine a horizontal standard that would be meaningful across applications and sectors” within the paragraph. See the paragraph below with the proposed changes: “Using agreed standards for design, manufacturing and business practices can function as a quality management system for AI offering consumers, actors and governments the ability to recognise and

disabilities, and how was this designed into the system and how is it verified?• What definition(s) of fairness is (are) applicable in the context of the system being developed and/or deployed?• For each measure of fairness applicable, how is it measured and assured?• Was a diversity and inclusiveness policy considered in relation to recruitment and retention of staff working on AI to ensure diversity of background?• What is the purpose of the system and who are its target users?On point 4 “Governing AI autonomy”:We suggest to amend the first two bullet points to add precision and to delete the fourth bullet point on “the overall responsibility of human beings”, as it seems to merely repeat the first bullet point. We also suggest to amend the last bullet point to take into account a situation where a developer builds the tool but does not use it. See the final list of bullet points below with the proposed changes:• Is a process foreseen to allow human control, if needed, in the relevant stages?• Is a “stop button” foreseen in case of self-learning AI approaches? In case of prescriptive (autonomous decision making) AI approaches? What is the procedure to make use of such button?• In what ways might the AI system be regarded as autonomous in the sense that it does not rely on human oversight or control?• What measures are taken to audit and remedy issues related to governing AI autonomy?• Within the organisation who is responsible for verifying that AI systems are properly developed and governed?On point 6 “Respect for Privacy”:We suggest to amend the last two bullet points, according to our feedback in Chapter II. See the last two bullet points below with the proposed changes: • How can users seek information about the use of their data? • Is it clear and is it clearly communicated, to whom or to what group issues related to privacy violation can be raised? On point 7 “Respect for (& Enhancement of) Human Autonomy”:We suggest to delete the first and fourth bullet points, to amend the third bullet point and to introduce a final bullet point on the change in users preferences. These changes ensure that the questions are more meaningful in allowing people to realize their autonomy. See the final list of bullet points below with the proposed changes:• Is useful and necessary information provided to the user of the service/product to enable the latter to take a decision in full self-determination?• Does the AI system indicate to users that a decision, content, advice, or outcome, is the result of an algorithmic decision of any kind? Does the particular use case require such information, and at what level of detail?• Are there mechanisms in place to allow users to communicate a change in preferences or provide feedback on the accuracy of algorithmic decisions? How is such feedback processed internally?On point 10 “Transparency”:In “Purpose”, we suggest to amend the third bullet point as follows: “Have the limitations of the product been specified to its users? Does the user case warrant this information, and to which level of detail?”In “Traceability”, we suggest to delete “On the reasons/criteria behind outcomes of the products” from the first bullet point, as this is unclear. We also suggest to delete the second bullet points on “the nature of the product or technology” as this repeats parts of the questions of the

case". See the paragraph below with the proposed changes:"[…] If one is a consumer or user of an AI system this entails a right to decide to be subject to direct or indirect AI decision making, a right to knowledge of direct or indirect interaction with AI systems, a right to opt out and a right of withdrawal according to the use case." On the "principle of justice: be fair", the text states that "justice also means that AI systems must provide users with effective redress if harms occurs". Additional clarity on the purpose of this provision would be welcomed. Today, if a government decision is wrongly taken based on AI systems, the decision can indeed be appealed like any other government decision. On "the principle of explicability: operate transparently", the Guidelines unfortunately propose some impracticable measures which would put a limit to innovation and allow only simplistic systems. For instance, no other economic field requires "business model transparency", which would be impossible for companies to provide as business models change over time. Such transparency is provided by terms of service and the GDPR. We suggest to delete "both technological and business model transparency matter from an ethical standpoint" and "business model transparency means that human beings are knowingly informed of the intention of developers and technology implementers of AI systems." We also suggest to add "depending on the application. It does not imply disclosure of source code or any other information that would threaten industrial property or trade secrets" after "levels of comprehension and expertise". See the paragraph below with the proposed changes:"Transparency is key to building and maintaining citizen's trust in the developers of AI systems and AI systems themselves. Technological transparency implies that AI systems be auditable, comprehensible and intelligible by human beings at varying levels of comprehension and expertise depending on the application. It does not imply disclosure of source code or any other information that would threaten industrial property or trade secrets".In addition, the paragraph on explicability mixes general principles with AI-specific issues by using "informed consent". For example and as explained above, "informed consent" has a specific meaning according to the GDPR, narrower than is probably meant here. We suggest to replace "informed consent" by "trust", which is the focus of these Guidelines. And as it may not technically be possible for all individuals and groups to "request evidence of the baseline parameters […], we suggest to at least nuance this approach by adding "in certain cases and based on agreed procedures". See the paragraph below with the proposed changes:"Explicability is a precondition for achieving trust. Explicability also requires accountability measures be put in place. In certain cases and based on agreed procedures, individuals and groups may request evidence of the baseline parameters and instructions given as inputs for AI decision making (the discovery or prediction sought by an AI system or the factors involved in the discovery or prediction made) by the organisations and/or developers of an AI system, the technology implementers, or another party in the supply chain." On point 5 "Critical Concerns Raised by AI"Generally reward ethical conduct through their purchasing decisions. However, the nature of AI makes it difficult to imagine a horizontal standard that would be meaningful across applications and sectors […]". On Education and Awareness to foster an Ethical Mind-Set (p22), we suggest to mention as well "public authorities" in the list of users of AI (as already mentioned above in the Guidelines) as well as "organisations overseeing the application" of AI systems. See the paragraph below with the proposed changes:"[…] Education here refers to the people making the products (the designers and developers), the users (companies, public authorities or individuals), the organisations overseeing application and other impacted groups […]". We also suggest to add the following new paragraph on Global Governance after the paragraph on "diversity and inclusive design teams" (p22): "AI and technology as a whole are often built and applied across borders. They are part of an ecosystem, where different components might stem from different regions in the world. For this reason we must maintain a dialogue with other geographies when it comes to the responsible development of AI. The most reliable way for Europe to ensure trustworthy AI for its citizens is to collaborate to promote a shared understanding and common norms across geographies. Europe should not miss the opportunity to shape the global debate on AI governance".

"Purpose" section above.

					speaking, while the issue of non-discrimination is raised in the next chapter, it would have been interesting to raise it in this section as well.On 5.1 (identification without consent), we suggest a clearer, more nuanced perspective by adding "one must also be mindful of what practices are harmful and not harmful, lawful and unlawful. Not all identification processes create a danger for the individual and many are actually beneficial" in the first paragraph; by adding "one should also consider that AI might make it easier to derive personal information from" in the second paragraph and by removing the first two sentences and parts of the last sentence of the second paragraph. See the paragraph below with the proposed changes:"[…] Differentiating between the identification of an individual vs the tracing and tracking of an individual, and between targeted surveillance and mass surveillance, will be crucial for the achievement of Trustworthy AI. One must also be mindful of what practices are harmful and not harmful, lawful and unlawful. Not all identification processes create a danger for the individual and many are actually beneficial. In this regard, Article 6 of GDPR can be recalled, which provides that processing of data shall only be lawful if it has a valid legal basis. […] Where the application of such technologies is not clearly warranted by existing law or the protection of core values, automatic identification raises strong concerns of both legal and ethical nature, with the default assumption being that consent to identification has not been given. One should also consider that AI might make it easier to derive personal information from "anonymous" personal data". On 5.5 (potential longer-term concerns), we strongly recommend the entire deletion of this section as it is highly speculative and does not add anything to the discussion. The Guidelines should focus on the current state of the technology to be practical and immediately applicable, and the principles included are broad enough to inform decisions on scenarios not yet foreseen.		
monica	gabrielli	Sogei - Società Generale d'Informatica S.pA.				2. Data governance: § What data governance regulation and legislation are applicable to the AI system? Have they been fully included in the design? Have the rationale for which certain standards may not be considered applicable or applied been described? 8. Robustness: Resilience to Attack: § What systems are in place to ensure data security and integrity or recovery? Fall-back plan: § Have fall-back plans been defined and tested? Are foreseen clear ways to inform the user of the execution of a fall-back plan?	We would like to propose, in the description of the development processes - includind analysis, design, development and use (ref. fig. 3 on page 19) the inclusion of an "Ethical test" to evaluate the comformity of an IA artefact to be adopted in each phase of the life-cycle of an artefact. Such a test should be included in the phase "analysis" and periodically performed based on the reles governing the business process of the IA artefact itself.

Anonymous	Anonymous	Anonymous	zu "Trustworthy AI"Den Begriff halte ich aus einer "ethischen" Sicht problematisch. Hier wird "Vertrauen" instrumentalisiert für einen einseitigen ökonomischen Nutzen: Wettbewerbsvorteil. In erster Linie müssen m.E. die Akteure, die AI für ihre Zwecke einsetzen, "vertrauenswürdig" sein. Sind deren Motive und Ziele des AI-Einsatzes integer? Werden diese Ziele ehrlich kommuniziert? Sind die mit dem Einsatz von KI verbundenen Versprechen der jeweiligen Akteure nachprüfbar?Zu "wellbeing als goal"/utilitaristischer EthikansatzAußerdem: Der hinter dem Papier stehende utilitaristische Ansatz ist sicherlich Mainstream und grundsätzlich begrüßenswert. Die Herausforderung ist aber, diese zu operationalisieren. Was wird unter "individual or collective wellbeing" konkret verstanden? Wird wellbeing auf die ökonomische Dimension reduziert (mein Eindruck) oder welche Komponenten des wellbeing spielen darüber hinaus eine Rolle? Gerade bei AI spielen m.E. auch Gesinnungs- und Verantwortungsethische Ansätze eine zentrale Rolle. Gerade der "Output" von KI Systemen kann in seinem utilitaristischen Nutzen aufgrund der oft vorhandenen Intransparenz und nicht eingeschränkten explainability autonomer ML-Entscheidungen nie vollständig beurteilt werden. Zudem sind die individuellen Präferenzstrukturen nie vollumfänglich zu erfassen und zu verstehen, so dass ein universalistisches "wellbeing" kaum erreichbar scheint .	zu 5.1.Diese Formulierung in Verbindung mit den Rechtfertigungstatbeständen zur Sammlung personenbezogener, anonymisierter oder pseudonomisierter Daten öffnet dem kommerziellen (Miss-)rauch von Identifizierungstools Tür und Tor. Es macht für eine Person keinen Unterschied, ob er auf Basis eindeutig identifizierender Daten z.B. per Bot als "Herr/Frau Müller" "manipuliert wird oder auf Basis von anonymisierten, aber das Individuum dennoch klar profilierenden Daten i.S.v. "als Herr/Frau Müller ähnlicher Mensch" identifizierbar.zu 5.2 auf S. 11"AI developers and deployers should therefore ensure that humans are made aware of – or able to request and validate the fact that – they interact with an AI identity."Das Problem aus meiner Sicht: Was ist eine „AI Identity"? Gemeint sind wohl (Chat-)Bots oder andere Tools zur automatisierten Kommunikation und Kundenbetreuung. Fällt da aber auch z.B. ein Algorithmus zur personalisierten Preisbildung darunter? Auch da müsste m.E. das Transparenzgebot gelten mit Blick auf: 1. Dass KI eingesetzt wird 2. Wozu KI eingesetzt wird (z.B. zur Preisstellung, Information, Service) 3. Auf welcher Datengrundlage KI eingesetzt wird (personenbezogene/anonymisierte/pseudonymisiert oder aggregierte/gruppenbezogene Daten)So deutlich formuliert finde ich das im Papier aber bisher nicht.zu 5.3. auf S. 12Dort werden das Thema „Citizen Scoring" und die damit verbundenen Risiken angesprochen. Es wird dort aber nur eine Bedrohung bei Verwendung solcher Scores durch „public authorities" gesehen. Scoringverfahren werden aber in vielfältiger Weise auch von anderen Akteuren eingesetzt. Z.B. von Unternehmen für kommerzielle Interessen, von Verbänden, Parteien und anderen Organisationen. Dies wird in 5.3. aber nicht thematisiert. Hier wäre unbedingt zu fordern, dass KI -Scoring-Verfahren grundsätzlich restriktiv, rechtlich kontrollierbar (Offenlegungspflichten) und allgemeinen Transparenzregeln gegenüber der Öffentlichkeit folgend anzuwenden wären, unabhängig von welchem Akteur auch immer. Ggf. müssen hier Verbraucherrechte gegenüber Wettbewerbsrecht und "Eigentumsrechten" der Algorithmeninhaber stärker gesichtet werden.zu 5.5.Nicht nur "Experten", Ingenieure und Wissenschaftler gilt es zu trainieren. Es muss Kompetenz bei allen Bürgern in der Verwendung von mit KI unterfütterten Frontends, Anwendungen, Kommunikationsprozessen aufgebaut werden.	zu II 1.Hier greift das Produkthaftungsrecht. Jeder, der ein System "in den Verkehr bringt", hat für die Folgen, die dadurch entstehen, zu haften. Es muss nachgewiesen werden, dass im Vorfeld eine ausführliche Risikoabschätzung erfolgt ist. Verantwortung bedeutet, Risiken im Vorfeld zu erkennen und zu minimieren.		Insgesamt ist mir der Entwurf zu Ökonomie lastig. Wellbeing ist mehr als wirtschaftliche Prosperität. Die Wettbewerbssituation mit Dina und den USA als Argument für den hinter diesem Prozess stehenden Zeitdruck zu verwenden, halte ich für nicht legitim, undemokratisch, den Prozess angehend schädlich und einer besseren Wettbewerbsposition nicht wirklich zuträglich. Manchmal gewinnt man mit der Follower Strategie. Wie die empirische Erfolgsfaktorenforschung zeigt sogar öfters, als mit einer Pionierstrategie. Goggle war/ist Follower (Web.de gabs vorher)! Facebook war/ist Follower! Den utilitaristischen Ansatz sollte man kritisch hinterfragen bzw. mit anderen Ansätzen anreichern!
Jean-Philippe	Steeger	CEC European Managers		The document defines the "ethical purpose" of AI as respecting the rights, principles and values as enshrined in the EU Treaties and in the Charter of Fundamental Rights of the European Union. Unfortunately, the delimitation between the concept of rights, principles and values appear rather vague and even tautological in their current formulation. The "rights-based approach" taken delivers insufficiently on an ethical case for these rights in proper terms. Furthermore, the document is ambiguous over the term "ethical purpose", since AI systems shall on the one hand "comply with" values, principles and rights (p. 3) and on the other serve them as a purpose. The	Remark on accountability: Complying with the human-centred approach CEC stands for, individuals shall remain at the heart of decision-making, the ultimate responsibility and liability for errors or biases in the system design shall lie in those in charge of the system. At all critical moments at least, a human evaluator and decision maker is needed, ideally with ethical knowledge and relevant skills. Remark on safety: to effectively ensure safety, another requirement is needed beforehand - the precautionary principle. The precautionary principle foresees that in the case activities can lead to morally	The list represents a helpful tool to assess AI systems.	The document is a starting point for an extensive reflection upon the ethical, social and economic implications AI has and could have in the future. Since the development of AI will be shaped jointly by management decisions, systems design and legal frameworks, the EU can now create a level playing field for implementing socially, economically and environmentally beneficial systems.

latter case implies that AI, and thus also organisations developing it, can only be ethical if they serve the purpose of advancing fundamental rights. At the same time, these rights and their underpinnings can evolve over time, making the need for a stronger ethical foundation of the guidelines even more important.

Central question: what shall we do?
Shifting away from the questions of rights, it may be argued that the ground-breaking trait of AI lies in its unmeasurable potential to create a utopian or dystopian society from the contemporary point of view and compared to previous technologies. This brings up classical ethical questions about the "good life", as well as the Kantian questions about what the human being is, what the human can hope for, what it can know and what it should do. Since the human, at least seemingly, could soon know and hope (for) almost everything, the central question appears to be: what, if almost everything is indeed possible, should the human do? And who is the human in this position? Of course, these questions are closely related to the purpose of work both conceptually and factually as a historically defining feature of human life.

Ethical principles and challenges
Later in the chapter, a set of five ethical principles is defined: beneficence, non-maleficence, autonomy, justice and explicability. Considering the powerful long-term potential of AI for delivering on some of the most pressing contemporary challenges, the principle of "sustainability" could be added. AI could have a positive long-term effect to ensure a living basis for everyone (cf. SDGs) and to limit pressure on Earth's life-supporting systems. On the other hand, the development of AI itself, in terms of energy and raw material use, has to be examined critically.
Finally, the last part of the chapter discusses some ethical challenges posed by AI, including consent, transparency of AI systems, mass citizens' scoring, lethal autonomous weapon systems and potential long-term concerns. As far as scoring systems are concerned, employees should be protected from extensive and unnecessary surveillance and have the right to be forgotten at the end of their employment relation. When it comes to the speculative long-term concerns, CEC reaffirms its opposition to a techno-deterministic view, which would acknowledge the possibility of artificial consciousness or attributing rights to technical objects performing tasks, even if complex and seemingly humanoid. This view is contrary to a human-centred approach to AI and stands in contrast to humanistic, religious and evolutionary worldviews. The EU should be clear that it is at the service of humans, not technology.

unacceptable harm, even if uncertain, measures should be taken to avoid or diminish it. "Morally unacceptable harm" usually refers to harm to humans or the environment that is: "threatening to human life or health, or serious and effectively irreversible, or inequitable to present or future generations, or imposed without adequate consideration of the human rights of those affected" . A sound risk analysis and its constant update are needed to assess the potential damages.

Remark on stakeholder and social dialogue: making use of the diversity among workers, managers and employers and other stakeholders can be a tool to flexibly adapt to developments of AI and labour market related implications. Social dialogue in particular can inform decision-making on the development of AI systems at company, sectoral, national and European level. Societies in the future will also require institutions to hold deeper and critical debates about the implication technology has on work and life. Ultimately, social dialogue may gain in importance to fit this requirement.

Remark on education and awareness to foster an ethical mind-set: throughout lifetime, prospective decision-makers and AI designers shall be equipped with the necessary knowledge and skills to deal with ethical questions. Being able to understand the technical, ethical and socio-economic implications of AI will prove increasingly important. Particularly a scenario of self-reinforcing algorithms, based on utility calculations, may prove both ethically and economically problematic – requiring critical and empathic humans. Such algorithms are ethically problematic, because they may – particularly if no measures for traceability are implemented – restrict the scope of potential decisions illegitimately, de-facto excluding contingent developments (e.g. showing results only based on previous decisions). They are economically problematic, because the necessary space for creativity that is needed to innovate, could be restricted through algorithms serving a utilitarian logic.

Claude	Kirchner	CERNA (http://cerna-ethics-allistene.org)	We provide general comments about the document in the ``General Comments'' section below. These comments generically induce suggestions to be taken into account of this introduction.	The text shall probably develop the fundamental tension between current values and the fundamental reasons that constantly put them under pressure: novelty, efficiency, desire for transgression, blurring of borders. This will enlighten many other forces at work, some of them stronger than the one currently developed like charity, benevolence, etc. For instance, the ``Ethical Purpose'' part does not develop the side effects or effects not directly targeted. Traffic jams were not anticipated by the designers of the first cars. "What else is it going to do?" is a question that society asks today in every debate on new technologies. In particular, due to the complexities of the Human and IA systems (of the human being on the one side, and of AI systems on the other side? on human-AI systems?), uncertainty is an issue that should be better taken into account : let us mention that syntactically, the word "uncertainty" is mentioned only once in the text. About section 5 The theme of lethal uses of AI systems (5.4) focusing on LAWS (which EU countries characterise as "fully autonomous weapon systems") shall emphasize that the 2018 news (Maven Project with Google/US Air Force collaboration) have in particular shown that AI technologies for lethal actions are already being used today. In order to gain on generality, paragraph 5.4 could be extended to deal with "The uses of AI as tools of physical coercion", which includes both the uses of AI as "personalised" targeting tools (facial recognition), the prospective subject of LAWS but also, for example, all coercive border control technologies that FRONTEX may have to evaluate. This will allow to naturally introducing a new point about "The use of AI as a tool for psychological coercion". This shall allow highlighting the main concern for European democracies, which relates to the "use of AI as a tool of psychological constraint" (although the question of "potentially manipulative nudging" is mentioned on p. 17 and p. 26). Here again, the news gives specific examples (Cambridge Analytica scandal) of the immediate consequences affecting the European democratic model and all the ethical principles highlighted by the report (point 4, p. 8-10). The issue of manipulation for commercial and mercantile purposes is indirectly addressed in the section on "hidden AI systems" (5.2. Covert AI systems, p. 11): this should certainly be extended to manipulation carried out with a political or ideological objective.	About Data governance: the text will benefit from better hindsight on statistics and should mention data encryption. When discussing the "by design" (page 19), notice that implementing moral values or ethical principles does not warrant that the program will be legal, moral or ethical. Indeed, what is programmed is a model, i.e., an abstract representation, which can be very far (consciously or unconsciously) from the initial concepts considered. About technical methods 2. Technical and Non-Technical Methods to achieve Trustworthy AI : Traceability & Auditability In order to tackle the challenges of transparency and explicability and to help non-specialists or "laypersons" understanding the causality of the algorithmic decision-making process, the use of virtual serious games could be recommended.	This chapter is the main part of the document and we globally agree on the main guidelines provided, emphasizing again, as developed below in the ``General Comments'' section, that it should be clearly sustained by an European ethical reflection on AI.  An implementation project must accompany the guidelines. They should also be related to other main initiatives like the G7 connected ``IPCC of AI'' initially developed between Canada and France as well as the idea of developing an European AI Observatory. As things are evolving very rapidly with time and the speed of innovation developments, the Guidelines should be designed with a built-in dynamicity. In particular, the principle of the ethical expert (page 8) shall be included in the Guidelines of III.	It seems to us that the main theme of this document should be to establish a path from reflection on ethical issues in AI to a set of dynamic guidelines for all relevant stakeholders developing, deploying or using AI. Therefore Europe shall base its guidelines in terms of AI on a "European ethical reflection on AI" that is more foundational than a "Trustworthy AI made in Europe". Indeed, it is only through a dynamic and permanent ethical governance of AI that the trust ("trustworthy AI") sought will be inspired. The general development of the document could also take into consideration the following remarks: • The discourse of fundamental human rights is important and should take into consideration the intrinsic difficulty to specify human rights in a programming language. It is the designer who interprets them; but, in a democracy, the interpretation of the law is the responsibility of the courts. This leads to an inevitable dispute between the design of a technological product and its judgment by society. The document rightly states that "good intentions are not enough", and should talk about methods before drawing technical and in general partial "solutions". • The values, in particular those involving humans, are rarely fixed. Typically, speaking of human autonomy depends on culture, space and time. For instance human autonomy was not understood in the same way in 1019, 1819 or today as well as in Paris versus in deep wild forest. • A good example in taking into account these continuous evolutions is provided by the French law on bioethics that is by design revised every 7 years. • The informed consent is a challenge to be implemented fairly to the people or the institutions in the context of evolving, complex and learning algorithms where transparency and explicability could be difficult, today, to maintain. The case of consent in scientific researches should be specifically addressed. • Human responsibilities have to be categorized with respect to different functions. For instance in ``Research Ethics in Machine Learning'' (https://hal.archives-ouvertes.fr/hal-01724307), CERNA's recommendations identify specificities of programmers, trainers or users. • To design guidelines is important and the ones provided in chapter III are useful. They should be motivated but also challenged by the development of an ethical reflexion, in each situation and along the way. • Environmental considerations (not harming the planet and its resources) seem out of scope in the current document when in our opinion they are clearly part of the ethical issues to be considered. Indeed AI should be part of an "integral ecology" approach addressing global challenges for the 21st century and beyond.
Philipp	Ehmann	eco - Association of the Internet Industry	eco welcomes the work of the AI HLEG. Discussing the impacts of the use of (semi) autonomous systems in different scenarios as well as their implications is an important first step for the preparation of a more widespread use of technologies consisting of or making use of Artificial Intelligence. In	As set out before, the normative approach taken by the AI HLEG seems proper and favourable way forward. The chosen aspects of “Respect fo human dignity”, “Freedom of the individual”, “Respect for democracy, justice and the rule of law”, “Equality, non-discrimination and solidarity including the	The guidance given in this chapter is seen as useful for companies conducting impact assessments on and operating autonomous systems. The references made on the realisation of the factors relevant for trusted Artificial Intelligence seem plausible and legitimate. In addition and in alignment with		A rights based approach toward AI is from the current general point of view a favourable approach that, if applied appropriately, can actually proof to be an advantage for the European digital single market. Furthermore, a rational approach towards (semi) autonomous systems is

			order to have a successful take-up of modern AI-technologies both providers and developers and users and persons concerned of said technologies do need legal certainty for their activities and interactions with AI. The implications of autonomous system do represent a challenge not only for the general public, lawmakers and selected industry sectors but also for the digital economy and the internet industry. eco itself has published its "Guidelines for the Handling of Artificial Intelligence" <https://www.eco.de/wp-content/uploads/2019/01/20180918_eco_LTL-K%C3%BCnstliche-Intelligenz-EN.pdf> setting out fields of discussion for the implementation and acceptance of (semi) autonomous systems. A normative approach as it is taken through the AI HLEG seems the proper way forward when approaching the implications of (semi) autonomous systems, beginning with the most basic and encompassing fundamental rights of every individual. While this approach is generally favourable, the AI HLEG has drawn several conclusions that seem inconsistent with both the incentive on the lawful use of autonomous systems and the own governing standards set out in the document.	rights of persons belonging to minorities" and "Citizens Rights" all address factors relevant for setting up trustworthy AI systems. Covering them while setting up an AI is an understandable endeavour. When taking on the topic one should not forget that exploring AI and autonomous systems is still a challenge new to companies and so room for learning, testing and innovating systems should be available taking the before mentioned aspects into account. It should also take into account that (semi) autonomous systems are applied over different sectors and use-case scenarios, which are subject to specific governance, and regulation, which should also added to the canvas on the governing principles of said systems. Taking this into account ideas like implementing a general auditability and an AI-specific approach on regulation (set out on p. 10) might prove both dissuading for developers of autonomous systems and problematic for the implementation of said systems, which have to match cross-compliance challenges. This problematique is continued throughout the rest of the chapter when discussing general prohibitions on the use of autonomous systems and identification issues (both p. 11) and culminates in the debate on whether users or persons concerned should be informed whether they are interacting with said systems.	our "Guidelines for the Handling of AI", we also encourage a debate on existing factors and mechanics in economy and society, which require a broad debate on whether certain principles are regarded as acceptable. Often, it seems, it is not the technology, which needs to be governed but the very principles that define it. A wider approach, which also accounts for self-regulation, is a welcome addition for the uptake of (semi) autonomous systems, when all actors and parties involved have agreed on common principles.		encouraged and should envisage, as the document has set out, the different scenarios and contexts within which said systems are being deployed. A single-minded approach that decides ultimately on the rules, regulation and guidelines for (semi) autonomous systems is bound to fail. A dynamic approach on the technology and its use cases is encouraged and should also reflect the status of this paper as a living document as set out in the document. Aside from functionality, the acceptance of Artificial Intelligence should be bolstered through education of the public and conveyance of the functionalities of the systems deployed.
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Hala	ELROFAI	TNO	1- in the Executive Summary section, it is mentioned 'Importantly, these Guidelines are not intended as a substitute to any form of policymaking or regulation'. my comment: How they relate/support each other? It would help to give a short explanation in this document.	1- In the following section 4. Ethical Principles in the Context of AI and Correlating Values -->The Principle of Autonomy: "Preserve Human Agency" it is mentioned that: 'Human beings interacting with AI systems must keep full and effective self-determination over themselves. If one is a consumer or user of an AI system this entails a right to decide to be subject to direct or indirect AI decision making, a right to knowledge of direct or indirect interaction with AI systems, a right to opt out and a right of withdrawal' my comments: This make the introduction of L5 of automated driving vehicles 'almost' impossible. Since the driver is by definition is out of the loop and the self-driving vehicles interacts with other road users. Neither the driver or other road users are controlling the self driving vehicles. 2- in section 5. -->5.1 Identification without Consent my comment: This is limited by the nature/purpose of the application/situation. Hard to generalise. 3. in section 5. -->5.4. I find this application not Ethical at all. They should be a clear rules for which AI can applied.	1- in 1. Requirements of Trustworthy AI --> 2. Data Governance my comment: High Data quality is essential for machine learning solutions. Therefore they should be more implicit requirements and guidelines to ensure the proper/sufficient quality to ensure acceptable/good AI performance. 2- in 1. Requirements of Trustworthy AI --> 3. Design for all my comment: Hard to generalise. AI is a mean for achieving certain applications. This very much depends on the area/application that uses AI. This application could be subject to very limited/especial group of users. 3- in 1. Requirements of Trustworthy AI --> 4. Governance of AI Autonomy (Human oversight) my comment: This hinder/makes it difficult to the introduction of solutions/applications that interact surrounding environment including humans and have high level of automation. See my comment for Chapter 1, I have mentioned self-driving vehicle as an example. 4. in 2. Technical and Non-Technical Methods to achieve Trustworthy AI--> my comment: I miss one aspect that important for trust worthy AI: Controllability. Maybe it is mentioned to some extend by not explicitly. At TNO controllable AI is define as: AI-based systems should be designed so that humans, not computers and their algorithms, ultimately remain in control of, and thus morally responsible for, relevant decisions of AI system.	I am sorry that I did not have the time to read this section in details.	I have really enjoyed reading through this well written document. I suggest to have a summarised, 2-page, handout for the final/revised version. It will help to spread the message through, for awareness and therefore more feedback. The current document is rather extensive/long for many audience. Audience who are interested in more details they can always read the full version.

Anonymous	Anonymous	Anonymous		5.5 There is nothing to be done about black swans. Consideration should be taken for secondary consequences, most of which seem obvious to sociologists and anthropologists. A corporation may not be held responsible for social changes per se, but recompense should be done through taxation during the process of automation.	2. "and one has to be able to prune these [biases] away before engaging in training". This might be a bad idea, if not impossible. One can lower their importance through weighting. 3. Article 21 of the Charter of Fundamental Rights of the EU uses the word sex, not gender. (Gender is used again in sections 5 and 7). Some perceive these words to be synonymous and some do not. It would therefore be preferable to stick to the word from the legal document. 4. [Footnote 24] "Responsibility for behaviour lies with the developer" and "responsibility has to be with the developer". Yes and no. The developer might have implemented the behaviour, and the design might have been created by the architects. Responsibility lies with the programme manager, policy maker or software company owner/board. 8. Accuracy. "mitigate and correct unintended risks" => mitigate risks and correct flaws. Diversity and inclusive design teams It might certainly be necessary to have some kind of access to sources of information and for AI systems providers to have proven that they have done an impact assessment. It seems unreasonable to expect diversity in all teams, especially in small companies.		It would be good to define meaningful measurement solutions that would define an objective measurement system.
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Mila	Dimitrova			Section 5.5 is critical to adopting AI Ethics Guidelines that will be sustainable in the next decades. Specific attention should be put on the case of Artificial Moral Agents that might be combined with normative citizen scoring in a hypothesis to maintain rule of law in some totalitarian regimes. In addition, if EU companies invest in R&D to develop such AI technology it may be purchased from other countries outside EU jurisdiction and in this sense which might not have direct impact on EU citizens but will support unethical practices contrary to the UN Sustainable Development Goals. Restricting such technology at its genesis will be in common good worldwide as distribution will be more difficult to stop if production is ongoing.		Point 6 of the list might: additional consent to seek if the profile of the user is enriched and detailed profiling can be argued. The additional consent guarantees better grasp of the possible harms to privacy which are still difficult to be understood fully by data subjects as some rights are not material when it comes to data protection.	

Enrico	Nardelli	Informatics Europe	EXECUTIVE SUMMARY(1)The general tone appears to be techno-enthusiast and appears to take for granted that problems may be solved just by means of an appropriate technique while in reality political debate and consensus are required.For example:(1a)AI is key for addressing many of the grand challenges facing the world, ... ==>AI appears to be an important technical factor for addressing many of the grand challenges facing the world, ...(1b)on the whole, AI's benefits outweigh its risks ==>if properly managed, AI's benefits outweigh its risks(1c)since human beings will only be able to confidently and fully reap the benefits of AI if they can trust the technology ==>AI potential benefits will be confidently and fully reaped by human being only if they can trust the technology(1d)In subsection A. Rationale and foresight of the guidelines of the Executive Summary"Artificial intelligence helps improving our quality of life" ==>"Artificial intelligence may help improving our quality of life"(1e)In subsection A. Rationale and foresight of the guidelines of the Executive Summary"It helps optimizing our transportation infrastructure" ==>"It may help optimizing our transportation infrastructure"(2)THe discussion of the two components of "Trustworthy AI" should state "has to" and not just "should to". Trust requires a commitment and not just an effort to commit.(3)It should be made clear since their first appearance which are exactly the "fundamental rights" and "core principles" the "ethical purpose appeals to". This precision regarding which rights are addressed should also be present in the definition of "ethical purpose" in the Glossary. Exactly which right are addressed starts to be discussed only on page 3 (subsection B. A Framework for trustworthy AI) when it is said they are the ones "prescribed in the EU Treaties and in the Charter of Fundamental Rights of the European Union" while this should be clear since the very beginning. A more precise reference then appears only in the footnote of page 5 in Chapter 1, which is still not completely specified, since it says "These rights are FOR INSTANCE reflected in Articles 2 and 3 of the Treaty on European Union, and in the Charter of Fundamental Rights of the EU." Given that Article 2 of the Treaty is fully subsumed by what is in the Charter, while Article 3 is discussing both rights and goals and the rights are, again, subsumed by what is in the Charter, it is adviced to make reference just to the Charter.(4)The explanation of "ethical purpose" appearing in various points of the executive summary should explicitly include the provision to "Pay particular attention to situations involving more vulnerable groups such as children, persons with disabilities or minorities, or to situations with asymmetries of power or information, such as between employers and employees, or businesses and consumers." that appears in the box "Executive guidance" at page (ii). More specifically, the definition of "ethical purpose" in the glossary should be modified so as to include this provision.(5)The discussion of "technical robustness" in the executive summary should include a reference to "explicability". Moreover, a definition of "technical robustness" (including the need for explicability) should be added to the glossary.(6)In the Executive Guidance box,	(1)The discussion of "ethical purpose" should explicitly include the provision to "Pay particular attention to situations involving more vulnerable groups such as children, persons with disabilities or minorities, or to situations with asymmetries of power or information, such as between employers and employees, or businesses and consumers." that appears in the box "Executive guidance" at page (ii).(2)The discussion of principle of non maleficence lists among harms: "physical, psychological, financial or social". First of all is not clear why the term "financial" is used instead of the more general "economical", and then it appears peculiar that among all possible harms to the society, that fall within the category of "social harms" this one has been singled out and not anyone of the many possible other kinds of harms, e.g. "cultural" or "political".(3)The discussion of principle of non maleficence only characterize "vulnerable groups" demographically and not socially. Why, for example, a category of workers should not considered to be a "vulnerable group" who should have a place in the design process?	(1)It is not at all clear the relation among the 10 discussed requirements and the 5 principles discussed in the previous chapter. It is instead important to show how each requirement constitute the operationalization of (part of) one or more principle(2)(On Architectures for Trustworthy AI)AI itself can be successfully exploited for checking compliance of AI systems, with respect to norms and constraints imposed by some regulatory framework. To this end a two level, possibly hybrid architecture can be envisaged where an AI "supervisor", representing the regulatory component, could be used to guarantee the run-time compliance of a more opaque AI component by detecting deviations. (3)(On Testing & validating): The issue about "Verifiability" of AI systems could be introduced and discussed. The goal is to automatically verify that an AI system is provably correct with respect to some properties specified in a formal way. Verification techniques are applied in different domains in computer science. AI systems highly interact with the environment and evolve in order to adapt their behaviour to new situations. Therefore, determining formal specifications and proofs for AI systems is a challenging but very hard task.(4)(On Regulation)The challenge is how regulating AI systems and maintaining, at the same time a good balance with flexibility. (5)(On Education and awareness to foster an ethical mind-set): A question is how to address ethical issues in AI courses. AI courses and curricula have to guarantee that students will learn non only to be good computer scientists and practitioners, but also AI designers aware of the great impact that AI systems and tools have in the society. It would be useful to provide suggestions and guidelines for teachers about how to integrate AI ethics in AI courses.	No comment	Comments have been prepared by Informatics Europe with the help of Paola Mello, Univ. of Bologna (Italy), member of the AI Alliance.

			at page (iii)"Strive to facilitate the auditability of AI systems..." should be Require to facilitate the auditability of AI systems... and also"To the extent possible, design your system to enable tracing individual decisions..." should be "Require to design your system to enable tracing individual decisions..."(7)In the definition of "Human-Centric AI" in the glossary "strive to ensure" should be "require to ensure"(8)In subsection A. Rationale and foresight of the guidelines of the Executive Summary, top of page 2, shouldn't"Trust in AI includes: trust in the technology .... – or trust in the business and public governance models." be"Trust in AI includes: trust in the technology .... – AND trust in the business and public governance models."?			
Stephen	Pattison	Arm Ltd	The Working Document is an important contribution to the debate about how to ensure AI succeeds in Europe. Like other companies, Arm believes that AI will not succeed, i.e. it will not be widely adopted, unless it is trusted by the public. Arm too has been looking into how to build that trust, and hence into what ethical AI really means in practice.In general the ideas in the last two sections of the WD seem a good basis on which to build. Some of the ideas in the first section are more difficult. Before addressing some specific points in the paper, we would like to highlight a few key points:(i) What would Guidelines look like?The Working Document as a whole is obviously not yet in a form which would be easy for companies to use in an attempt to build trustworthy AI. Section one reads like an exploratory and discursive Academic piece of work. It covers a very wide range of topics, not all of which would necessarily be the immediate concern of business. This is perhaps inevitable at this early stage. But industry will need something more practical if Europe is to succeed in building trustworthy AI.Sections Two and, particularly, Three, contain important ideas for what might be included in guidelines. These need to be phrased in ways which clearly and simply describe the measures that companies should be prepared to take. The Guidelines should also aim at giving customers an easy-to-understand account of the measures companies have taken to	Section One of the WD seems sometimes to have some very broad concepts in mind: Business will need greater clarity if it is to put ideas into practice. It might be helpful to distinguish, as Arm has tried to do, between three different areas where ethics applies:(i) how do we want AI to work? (ii) how do we want to use AI? (iii) how can we manage the impact of AI? In group (i) how do we want AI to work, we can list many of the points in the WD: we want AI to be ethical by design, we want it to be transparent and explainable (or as the WD says, traceable and auditable to the extent possible) , we want to avoid both illegal and unfair bias, we want to know who is liable, we want human oversight for certain decisions etc.In group (ii) how do we want to use AI, we might include something about not causing harm, not interfering in national government or judicial processes, and about trying to control the use of lethal Autonomous weapons. More generally we might look here at whether we want to use AI for predictive analytics. Or whether we are only content for predictive analytics to be done in certain circumstances? In group (iii), how can we manage the impact of AI, we might want to talk about jobs, inclusivity etc.Breaking down AI Ethics into groups of issues on these lines can help focus on what different policy instruments might be used and by whom (business, Government, society) to address some of the issues. The WD describes an intellectual underpinning of	These sections are important and useful. They are a good catalogue of high level issues. They are naturally at this stage generic and high level. In some cases this avoids the actual difficult parts of the questions e.g. how to operationalise some of the principles. There is growing international agreement on the key problems around AI Ethics. There is less agreement on what companies should do about them. This includes determining what is right and wrong in specific of these situations is more difficult. It also needs to address how performance will it be audited and monitored and who will have responsibility and liability.	The Assessment list is a good place to start. In some cases more clarity is needed eg Section 7 'respect for human autonomy'. Would nudging a consumer to stay online violate this principle? Terms like 'full self determination' are not clear. Section 10 on Transparency tries to cover widely different notions of transparency. Some of it seems to require transparency over business models and 'limitations' of the product/service (which are not issues unique to the AI sector and should probably not be included). The points on traceability are important and are directly related to AI. It is important to remember though that any information made available can be provided in a way which consumers will easily understand.The ideas in section 8 on Robustness, and in particular Resilience to Attack, could be expanded. Many AI systems will be linked to other systems like IoT. We need to ensure that the whole system if resilient to attack. There are various templates available for checking Security processes and provisions. So we might add here 'have you used a recognised/third party template for providing security for parts of/the IoT parts of your system?' In some other cases it would also be helpful to have a more granular list of issues. For example one idea might be to look at whether it would be helpful to provide for all employees working on AI to undertake a course in AI Ethics? If so, who would provide such a course? Could it be managed remotely? More detail could be	A good start, particularly in the Assessment List.

						provide trustworthy AI systems. This may require two separate 'assessments': (i) as a company offering an AI service, what do you do to assure yourself that an AI system you are developing/using is trustworthy and (ii) how do you communicate that simply to customers.(ii) The need to encourage AI business in EuropeIt is wrong to think that there is a trade off between ethics and stimulating innovation. There will be no AI unless there is trust. Sometimes we may be uneasy about the results of AI analysis, and we may need to address that. But that is separate from the question of whether those results were fairly obtained, and properly explained, which is at the heart of AI Ethics.We need to ensure that Europe does not impose unnecessary obstacles in the way of developing AI businesses in Europe. This means we need guidelines which are clear and realistic.. Much of Section One of the Working document uses abstract concepts which appear very broad. (iii) The Commercial Uses of AI The document downplays the commercial uses of AI. It is strong on the health benefits. (And there is some vague language about it can help achieve wider public goals.) But unless we also take full account of the commercial scenarios we will not be able sensibly to address the ethical issues which might arise.Here we may have to acknowledge that personalisation of information and services has been a key characteristic of the growth of the digital business sector. The business model of many digital services may increasingly depend on their ability to personalise information, eg through targeted advertising. (iv) Is the regulatory landscape adequate for AI to succeed?GDPR has been a landmark in promoting the need for proper handling of personal data. But there has been some criticism of whether it can remain appropriate for an AI or Blockchain world. Simply put, as we get to full AI, we will have machines interrogating data in ways we haven't thought of: looking for patterns and linkages we have not thought to explore. So the notion of explicit informed consent for all aspects of data processing might be challenging. This is something we should be thinking about now.	some key ideas, tracing them to the concept of Human dignity and the Oviedo Convention. If we are to do this we need to be confident that the underpinning principles enjoy wide support. For example, not all the major European States appear to have ratified the Ovideo Convention.		provided in the section on how to handle complaints: are they dealt with by humans, within a certain time frame etcOn the question of bias it is important to distinguish illegal bias, which remains illegal whether done by AI process or not, from unfair bias – bias which is not necessarily illegal but seems 'unfair'. The later is much more difficult to define and guard against. In a sense many AI algorithms are 'biased': processing of travellers' data to identify potential smugglers is likely to start from some core characteristics of smugglers which might be said to be biased against those who are not smugglers but happen to fit the core characteristics. This is probably unavoidable. But at what point does it become unfair and undesirable?	
Richard	Krajčoviech	Independent Consultant				The guide is very hard to read. It is promising in chapter B three chapters, each offering a further level of abstraction, but because of missing cross-references and different terminology, it appear more than three more-less independent views on the human-centric design. E.g. the respect for democracy, justice and the rule of law (a header form fundamental rights) is not mentioned and hard to find in the ethical principles as likely as in the "realizing trustworthy AI". The idea of explaining the requirements through several level is great, but its implementation is far from perfect. When I am saying this (apologies for being so direct), I have on my mind a poor guy building his start-up and trying to dig though this document and apply it to his wise area. Or think of how much effort would you nead to teach this the university students. The other important comment is about distinguishing what is expected from designers and what is expected from the AI

system. Examples: "Good AI governance should include accountability mechanisms" does not help much to the startup guy without further explanation. I would expect clear statement that "(Considering current and foreseeable state of the art), the accountability for consequences of AI driven systems is shared among developers, producers, deployers and users (if informed properly), because AI systems are not able assess consequences of their actions in the scope required for being accountable. Accountability with AI systems would mean that we allow creation of machines that do harm (e.g. kill people) without anybody being responsible for their actions." Proponents of accountability with AI systems should think of their relatives being insulted by such machine and what they will do after such accident.

I am also missing better explanation of proportionality. Again, a startup guy building his clustering algorithm for recognition of some type of news does not know from this guide, what he should do to be compliant. He can have clue about data governance, privacy, robustness, safety, transparency and non-discrimination. How to make it accountable, designed for all, how to govern its autonomy and respect for human autonomy? Does he need Ethical AI review board? I think we can distinguish handful categories of AI systems based on predictability of associated risk:
1. Those which behave deterministically and can be properly tested, including their reliability (i.e. with risks similar to non AI systems)
2. Those with predefined set of actions and well controlled playground, so we can analyze typical scenarios and worst case scenario and their probabilities, even if they are not deterministic.
3. Those with predefined set of actions with non-deterministic behavior, but without specified playground, i.e. used in general public or facing unpredictable inputs with unpredictable reactions, physical or virtual. We do not know and cannot analyze well the associated risks and we need to take proper precautions.
4. Those allowed to invent new, unpredictable actions, but which can be used only in playground with well controlled boundaries. (Is this possible?)
5. Those allowed to invent new, unpredictable actions and used in general public.
6. Artificial Consciousness.

The guidance for types 1 and 2 is not much different from non-AI systems, except higher probability of unintended consequences, if not tested properly. They need application of extra rules in design and extra testing, but do not need any functionality to check, whether they are doing good or so - designers and/or deployers have still sufficient control over the possible harm. We should be careful with imposing too many requirements here, because then the guidance will be ignored by the business.

The guidance for 3, 4 and 5 might need, beside the intended functionality, additional features to ensure human centric behavior. We can be more demanding here, but still the requirements should be proportional to capabilities and autonomy of the systems.

						I have many specific comments, but I do not know, whether I manage to post them within the deadline (and I cannot blame the deadline at all). I will do my best, but for a case, I am submitting at least these general comments. I hope they are of help.
Anonymous	Anonymous	Anonymous	Simple and straightforward approach to AI Ethics. The construction of a framework for trustworthy AI, based on ethical core values and principles and applied in use cases, seems the best scenario for a European leading position in AI Ethics. Consistent and properly framed, the guideline structure suggest a good bases for trustworthy AI. No significant changes needed in this section.	The idea of making the core values and principles human-centric makes a lot of sense for the European AI-community. The ethical purpose by dynamic principles-values-rights depicted in Figure 2 suggests a significant reduction of uncertainties regarding active topics in AI, allowing the human-centric idea to be applicable. The five principles and correlated values described in the Chapter will be the proper base for the ongoing ethical purpose. No significant changes needed in sub-sections 1, 2, 3 and 4. Regarding 5. Critical concerns raised by AI: (Subjective observation) I personally feel most of the topics dealt with in subsection 5 lack the clarity of the full applied AI in the real world. All 5 examples (including the potential longer-term concerns) seem to be dealt and solved with a significant lack of knowledge and they all contradict directly or indirectly what has been proposed during the introduction and the first 4 subsections of Chapter I. Everything after this subsection (Chapter II, Chapter III and Conclusion) seem less important or significant, since the suggested concerns show that the High Level Expert Group is lacking communication with other groups and even their own understanding of their own draft. Major changes are needed (maybe even mentioning the concerns but not having such a reckless view) in order for the rest of the draft to keep its importance.	The mapping of the principles into the requirements of Trustworthy AI makes sense. But the realisation of trustworthy AI depicted in Figure 3 proves that the appplication of the techinical and non-technical methods to implement the requirements lack a proper evaluation and justification in an ongoing bases in the analysis-design-development-use circle. I suggest a better integration of the requirements, whether by defininining them in more depth or adding requirements that integrate the wholeness (general overview) of the system better; not just tackling characteristics seemingly independent one from another. This problem is reflected in the lack of completeness that can be read specially in the "non-technical methods" subsection. Significant changes needed in this section for sake of clarity and fully usability of the guideline.	Since the disclaimer of the assessment list as "preliminary only", my feedback in this chapter will be less objective and more subjective and general, regarding the topics dealt with. Since there is no real integration between the requirements (read my comment about Chapter II) the assessment list sounds extremely limited in its "assessment" quality. Maybe the addition of general (overall) requirements will make this list a bit more useful for the aimed target audience. Significant changes needed in this section for sake of usability of the guideline.	Besides the lack of clarity of the draft content and opinions lacking of proper knowledge in subsection 5 of Chapter I, the guide feels like a very fruitful work between experts. With the addition of proper general (overall) requirements in Chapter II a better Assessment List can be created for future use of the guideline for the development of Trustworthy AI.
Attila	Soltész	KIBEV			Besides of code of conduct and standardization, I recommend to consider to incorporate the concept of certification, also certification bodies, in order to testify the trustworthiness. Similar to, or along with the section 5 of GDPR. It would enhance the trust towards these products more than just to be claimed (by manufacturer) to be trustworthy.	

Mika	Tuuliainen	Confederation of Finnish Industries EK	No further comments.	Ethical Principles in the Context of AIpage 9: The Principle of Autonomy. "Preserve Human Agency"The wording of the principle of autonomy is too far-reaching. Already today consumers, workers and other users are subject to automated decision making, whether based on AI or simpler applications. E.g. the right to opt out could conflict with existing employee obligations and lead to dismissals. The principle of autonomy should be limited to freedom from coercion, not all kinds of subordination.page 11: Critical concerns raised by AI. Identification without explicit consent is an existing, widely used practice not necessarily dependent on AI applications. The draft refers to GDPR article 6, which lists several legal bases for processing personal information. Picking out consent as the primary justification is unfounded and goes against the technology neutral approach of data protection. Identification and (other) processing of personal data should be allowed on any lawful grounds recognised by GDPR or other relevant legislation.	The ambition of the chapter is good in describing which areas to think through when working with AI. However, from the perspective of business, it should be more concrete. Since we asked from our member companies to give feedback on the content of the ten requirements of trustworthy AI (as well as on the whole guide), we did not receive any. This might indicate that the content is yet too abstract for the companies.	The idea of having a concrete assessment list is good and helpful. The proposed list fits its purpose well. As mentioned on page 28 in the note, the practical operationalization of the assessment list is more than welcome. The four user cases of AI and tailored lists in each of mentioned contexts will give value added for the reader. The Guidelines are quite general and high level but when the content gets more concrete it could be practically applicable on how to act ethically.	EK supports this draft AI document which is relevant, challenging and much needed. In addition to its far-reaching social impacts, the ethical viewpoints associated with artificial intelligence should also be taken seriously, as they will directly influence companies' business.The Guidelines for trustworthy AI will strengthen the uptake of AI, but from the perspective of business it should be more concrete. To make the Guidelines truly beneficial for the developers and users of new AI solutions, business needs a more actionable guidance with concrete mechanisms and best practices. Concrete advice/best practices are needed by developers to instruct them on how to act and what kind of factors to consider.It is also positive that the Ethics Guidelines of AI most importantly focus at providing guidance on how to implement these principles. Further regulation on AI and ethics would be premature and may create unintended problems and limit the business potentials. There are new technologies emerging and all frameworks should be technology neutral as far as possible to not hamper competitiveness and add regulatory burden on companies.The topic worth mentioning is also ethics of data collection. Artificial intelligence cannot be ethical if the raw material it collects, data, is not ethical. The same concerns data collected for developing AI, such as machine learning applications. The draft strictly focuses on the reliable use of artificial intelligence but does not include a review of the actual collection of data and its ethics, which are essentially connected to it. Before data can be refined using analytics tools, it must first be collected, organised, edited and stored. All in all, it would be essential to assess the impact of how the approach with ethical purpose will affect innovation in Europe, especially the competitiveness of companies focusing on AI. It should be remembered that the industry does not want or need extra barriers to business, which is one of the cornerstones of European welfare. EU should actively promote international research cooperation concerning the ethical viewpoints of artificial intelligence, information exchanges relevant to this theme and the mainstreaming of good practices. It should also serve as the pioneer of ethical discussion in the implementation of artificial intelligence initiatives across Europe.
Birte	Dedden	UNI Europa	- UNI Europa ICTS welcomes the possibility to contribute to the stakeholders' consultation and underlines the importance of a broad public debate and information on AI. This debate must result in clear ethical and social guidelines and standards with the aim of improving the living and working conditions of European citizens. - We acknowledge the innovative potential of AI and new technologies that can be beneficial for our society. However, these new technologies also create challenges and we are concerned about the possible risks and consequences relating to working conditions, skills and training, ethics, equality, health and safety (among others). Therefore, UNI Europa would like to underline the importance of addressing AI technologies and robotization as topics for collective bargaining at all levels (company,	- UNI Europa supports the human-centric or human-in-command approach suggested in the guidelines. We agree that it is necessary that humans always need to remain in control of technology and machines. Likewise, we agree that the use of AI needs to respect European values and fundamental rights. - We recommend an expansion of 3.2 "In an AI context, freedom of the individual requires protection from direct or indirect coercion, surveillance, deception or manipulation." Much profiling that lies at the heart of AI systems relies on a degree of manipulation of data. This is not least relevant in relation to the world of work and especially in the use of AI in Human Resources. - We welcome that the HLEG understands the need to ensure that those involved in the	- In order to achieve "trustworthy AI", we need to establish public, independent and autonomous organisations that can control and audit (labour) algorithms (e.g. to identify underlying biases and the objectivity of data sets that train algorithms). Likewise, the implementation of the ethical guidelines on AI must be monitored. A European observatory focusing on the ethics in AI systems could play the role of an independent watchdog, including in business. - We would like the advice „to always keep record of the data that is fed to the AI systems" from the heading of data governance included under Accountability. For workers, it is paramount that the datasets used to evaluate performance, or in hiring or firing processes is transparent and can be accounted for.	- We welcome that processes shall be examined in order "to allow a human control, if needed" (assessment list – governing AI autonomy). In so doing it should not be a question "to keep a human in the loop". We need clearly defined measures, that empower people to exert this control in all processes – regarding resources (technical equipment etc), organisational needs (time, liability, etc.) and qualification. - We suggest extending the list on the assessment of use cases (p.28) and add the question of processes, in order to use AI to ensure decent work (development and impact assessment).	- UNI Europa ICTS welcomes the call for Accountability Governance on page 21. The establishment of Data/AI Governance Councils in companies will indeed strengthen the accountability of AI systems and will address a weakness in the GDPR. The Council should consist of shop stewards and management and be responsible for holding management accountable and transparent to the use of AI and data. Whistleblowers should be able to address concerns to the Council and mandate the council to investigate on reported issues. - We welcome the process of developing guidelines for a trustworthy AI made in Europe, which encompasses corresponding "guidelines made in Europe", but would like to raise the question why non-European companies such as Google were granted full membership and full participatory right in

national and European). AI and robotics have a huge impact on the future labour market, as jobs will sometimes disappear or be transformed and other jobs will be created. We need to accompany this process and address the question of skills and training for the future workforce: need to ensure that training on necessary digital skills is provided by education institutions and companies, and that it is not the sole responsibility of the worker to keep up with the rapid technological developments. Employability needs to be promoted through upskilling and reskilling schemes for workers. Investment in formal, informal and life-long learning is key; we must enable people to work with AI or invest in competences that AI will not cover. It is important to develop action plans at EU and national level together with education providers and social partners in order to modernize education and vocational training. We therefore welcome the call from the ILO Global Commission on the Future of Work for "a formal recognition of a universal entitlement to lifelong learning and the establishment of an effective lifelong learning system. (https://www.ilo.org/global/topics/future-of-work/publications/WCMS_662410/lang--en/index.htm)
- The social partners play a key role in this and the EU should cooperate with them and national governments in order to identify which job sectors will be affected by AI. We need to understand the timeline and extent of changes in the labour market. The involvement of social partners is a must to find appropriate and future-proof solutions to concerns relating to employment, training, the nature of work, (in)equality or social systems and collective bargaining, especially at sectoral level.
- As AI and automation have the potential to transform not only simple tasks but very complex processes, we need to have a large public discussion about the areas in which the use of AI is reasonable and beneficial for society. Part of the debate should be the question of how the profits generated by AI should be re-invested for the common good by creating employment in domains such as care, health services, education or mobility. Employees should participate in the distribution of profits, e.g. through wage increases or reduction of working time. Moreover, AI wins should be used to strengthen social security systems. This could be a measure to address the problem of future job losses and the precarisation of employment relations in a-typical work (e.g. platform work) due to AI and automation.
- It is therefore important to integrate the aspect of the quality of jobs, decent work and social progress into the ethical approach in order to create a balance with the purely economic objectives underlying the creation and use of AI and robotics.
- The Human-centric approach (HCD) not only presupposes information, transparency, participation and traceability, but also requires specific negotiation processes regarding decision-making in view of the aims and implementation of AI-systems at a very early stage for stakeholders such as employees and their codetermination bodies. development and marketing of AI (researchers, engineers, designers etc.) act in accordance with ethic and social responsibility criteria. This should be addressed by changing educational priorities for technical subjects and by providing lifelong learning opportunities (e.g. by incorporating ethics and the humanities into training courses in engineering).
- Organisations and companies should develop tools to facilitate ethical discussions and decision-making throughout the whole design process. This should be completed by internal training programs on ethics for all employees. Such training should help employees understand the AI systems themselves, their rights in relation to said systems and their possibilities of redress, complaint etc.
- AI should provide an opportunity for workers to apply their skills and competences to the fullest while at the same time remain owner of the production process. This includes the principle of transparency in the use of AI systems in HR, like the hiring of employees or the performance assessment of staff. It is important to safeguard the rights and freedoms of employees in line with non-discrimination rules as regards the processing of workers' data.
- UNI Europa welcomes 5.1 – 5.4. We support that these examples raise real-life concerns of the adverse consequences of AI systems.
- In 5.2. UNI Europa urges the group to expand on the issue of the human's right to know they are interacting with an AI identify. This could be done through a "labelling" system. For example, online bots should be labelled as such. Users should be made aware of the use of bots and AI in customer call-centre or help desks etc.
- We would welcome that the employer-employee, employer-worker relation is explicitly mentioned in 5.3 as an example of power asymmetry.
- Taking into account the power asymmetry in employer-worker relations, a separate point 3.6 on "workers' rights" should be added, which should contain the following points: "decent work by design", equal negotiation processes in the sense codetermination rights, informational self-determination of employees, non-discrimination principle and freedom of association including the right to strike. This is needed in order to secure worker's rights to co-decide on aims and application of AI systems, and create a legal framework.
- Concerning the long-term risks and concerns we welcome that these should be considered. This could become an integral part of the accountability and audibility demands – i.e. that developers, users deployers etc need to reflect on the development/changing nature of the adopted AI as well as engage in predictions/forecasts of its future development scope and the consequences (positive and negative) hereof).
- AI's influence does not only affect the world of work, but also democracy and society as a whole. We welcome that the draft refers to this point in Chapter I, paragraph 5.3., by stating that AI is not to be implemented in order to enable "citizen scoring" by a state/government. But this should also apply to private businesses. Neither states nor companies should be
- The explanation of the principle of autonomy covers the question of AI at work only in a footnote, whereas this is an important issue that should be given a more prominent place. We would like to highlight the right of workers to individually and collectively opt out or withdraw from the use of AI systems (or a decision chosen by an AI system) if they undermine the workers' autonomy, decision making competence or disrespect fundamental rights and ethical principles. We recommend the inclusion of a special chapter that provides for ethical guidelines on AI in the work environment to address these issues more in detail.
- UNI Europa ICTS welcomes that the HLEG on AI acknowledges the importance of social dialogue to realise trustworthy AI. We would like to add that the involvement of social partners, and in particular employee representatives, should not only take place regarding the general public debate on AI. Social partners should be involved in the establishment of codes of conducts, of standardisation schemes, development of training and in the proposed accountability governance. Employee participation and inclusion should take place early in the design, development and deployment of new technologies including AI and robotics. It is essential and important not only to inform and consult workers representatives in the work place or at branch level, but to enhance their co-determination rights and ensure their right to co-decide on the aims, reasons and implementation of AI at the workplace.
- Social partners at all levels should be involved in the implementation at company, industry, national or international level, including through collective agreements setting standards. In this context, it is required to describe the negotiation processes, e.g. central control structures for sector-specific solutions (cf. ‚AI Now' Report 2018): „Governments need to regulate AI by expanding the powers of sector-specific agencies to oversee, audit, and monitor these technologies by domain." (https://ainowinstitute.org/AI_Now_2018_Report.pdf)
- Regarding the principles of accountability and transparency, we need to establish mechanisms for the protection of whistle-blowers who disclose the risks of AI systems or the non-respect of ethical principles – especially in the case of employees in companies that develop such systems. Internal reporting of risks and violations should be supported and rules in place to ensure follow up.
- Organisations and companies should pay attention to potential biases encoded in the system development, training data and model performance – especially those that my affect the most vulnerable. They could also establish an internal ethical review process to democratise the decision-making process
- Companies should not only increase transparency regarding the design and development of AI systems, but also in organisational chains of responsibility.

the High- Level Expert Group. The status of associate expert would be more appropriate.

UNI Europa also supports the position of the ETUC regarding this consultation.

				allowed nor have the possibility, to create human profiles such as “moral personality” or “ethical integrity”. We reject the proposed opt-out-function and even possible “opt-in”-functions are not to be designed in a way that they conflict with fundamental human rights and possibly lead to the waiving of services that are useful for a person. AI-based services, that are important for work and life, must be designed in such a way that they do not require the collection of data which could be useable for human profiling. - Creating big data-bases always includes the risk of hackability as well as intentional and unintentional data-leaks. The guiding principle of “data-sovereignty” needs data-security in order to be viable. This implies explicitly not surveying data in areas that are of highly explosive nature for people in e.g. political, private or work-related areas. Fundamental rights as informational self-determination, the freedom of association and freedom of speech are not to be put at risk by creating such data-bases.			
Sibylle	Gabler	DIN e. V.	We welcome the goal of the document to capitalize on opportunities of AI, while at the same time minimizing risks through clear guidelines. Standardization provides an excellent platform for this type of work, as it is already accessible to all interested parties. A uniform understanding of AI and ethical aspects is essential for the further development of the technology. This is only possible with standards that form a basis for AI products and applications at European - and especially international - level. They function as a basis for cooperation, even amongst members of different nationalities and areas of expertise. Hence ISO/IEC JTC 1/SC 42 works on a standard for "Artificial Intelligence Concepts and Terminology" and a " Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)". The active participation of European experts in European and international standardization must therefore be promoted so they can help shape AI standardization from the very beginning, taking European interests into account. EU member states and the institutions of the EU must ensure that representatives of the public sector fulfil their obligation to actively take part in standardization.	The document tables AI as a new technology (p. 7) but AI has been used and approved in healthcare for more than ten years. It may be more appropriate to explain why it needs another guidance at this moment.	In terms of mapping abstract principles into concrete requirements, standardization plays an important role. We therefore welcome that the draft for Ethic Guidelines for Trustworthy AI mentions standards as one method to implement trustworthy AI. We recommend to rewrite the paragraph on standardization (p. 21) as followed: “Using agreed standards for design, manufacturing and business practices can function as a quality management system for AI throughout the entire production and supply chain offering consumers, actors and governments the ability to recognise and reward ethical conduct through their purchasing decisions. Throughout the entire development process and after market launch, they create comparability, enable interoperability and contribute to the safety of products and processes. One example of such a standard is the DIN SPEC 92001-1 , that supports managing AI systems during the entire lifecycle and assures safe and aware development and maintenance of AI products and processes. Many regulatory framework require assessment of change in safety and performance which standards can address / aid to achieve. As standards are subject to regular review, continuous adaptation to the dynamic environment in which trustworthy AI operates is assured. Beyond conventional standards, co-regulatory approaches exist: accreditation systems, professional codes of ethics or standards for fundamental rights compliant design. Examples are ISO, CEN and CENELEC Standards, the Fair Trade mark or Made in Europe label.”	Standards are the most acceptable and widely used way to describe the state of the art in a respective field and operationalize superordinate concepts. If the trustworthiness of AI is to be assessed, compliance with standards is therefore an important indicator that should be mentioned in the Ethic Guidelines. An assessment of the trustworthiness of an AI-based system interfacing directly or indirectly with humans should always build on existing standards. An assessment list could therefore ask: • Has the AI-based system been developed according to the state of the art? • Were relevant standards taken into account in the development of the AI-based system? • Is the AI-based system compliant with these standards? Standards may further help to assess the safety of trustworthy AI (see p. 27).	We highly recommend a regular and structured exchange between the High-level Expert Group on AI and European and international standardization bodies (CEN, CENELEC, ISO, IEC) and their standardization committees on AI (e.g. ISO/IEC JTC 1/SC 42) as part of the further development of the Ethic Guidelines. Knowledge from current standards and ongoing standardization projects (including work of ISO/IEC JTC 1/SC 42/WG 3, the international standardization working group on trustworthiness) should be included in the guidelines. At the same time, findings from the Ethic Guidelines (e.g. the assessment list for trustworthy AI) might be transferred into current or future standardization projects.
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential

Anonymous	Anonymous	Anonymous	Page 2: A mechanism will be put in place that enables all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis. This will be set out in the final version of the document. This is a very weak endorsement and underlines the voluntariness of the support to the guidelines. But the document contains enough material as to guide policies and ultimately legislative process (i.e., laws). Consequently, current voluntary may become mandatory in the short run. Page 4: in the figure, it reads "To be continuously evaluated, addressed and assessed". Continuous evaluation leads to infinite loops. A more sensible approach would be to establish a limited review and correction process. Additionally, it is not clear who are the parties involved in evaluation (specifically, to provide contrast). Current practice in similar realms doesn't provide evidence on stakeholders' participation (for example, cybersecurity).	Page 7: It does not only entail freedom from sovereign intrusion, but also requires intervention from government and non-governmental organizations to ensure that individuals or minorities benefit from equal opportunities. Although it is not legally protected per se, new technologies promotes the clustering of individuals in more groups than only individuals or minorities. By the use of data correlation, individuals can be assigned to families (in the broader sense, including pets and appliances), lifestyle groups, etc. The protection should be also guaranteed for any voluntarily adhered group. Page 12: LAWS can operate without meaningful human control over the critical functions of selecting and attacking individual targets. Ultimately, human beings are, and must remain, responsible and accountable for all casualties. The topic of human accountability appears only for the case of lethal autonomous weapon systems while in the rest of the text the mentions to accountability are quite blurred. In general, any reference to accountability should be associated either to a natural person or to a legal entity.	Page 14: Good AI governance should include accountability mechanisms, which could be very diverse in choice depending on the goals. Following the above commentary, this item refers generically only to pecuniary compensation but not to the legal or penal responsibility. Page 18: Transparency concerns the reduction of information asymmetry. It is quite significant that transparency appears as the last requirement in the list when it may be apparent that it is the main factor for AI trustiness. The topic is insufficiently described and deserves much more depth in the definition of some key terms such as information (a)symmetry, explainability, data dynamics, morally significant impact. All of them looks like very common terms but are far from common understanding in the field of AI. Page 19: This also entails a responsibility for companies to identify from the very beginning the ethical impact that an AI system can have, and the ethical and legal rules that the system should comply with. Different "by-design" concepts are already widely used, two examples of which are Privacy-by-design or Security-by-design. Although the chapter is titled "Technical methods", most of the references are generic claims or desiderata. Ironically, the claims to the "by-design" concepts are paradigmatic to the case, as they are objectively that, i.e., concepts, with a very difficult progress in methodological content or practical tools. In the rest of the methods, there are no proper methods, but loose references to what it should be (without saying how). Perhaps, a list of recommended models, methods, and tools in an annex would result of better utility vis-à-vis actual implementations. The EC has funded plenty of projects aiming at developing those kind of components.		Page 29: This document forms part of a vision that emphasises human-centric artificial intelligence which will enable Europe to become a globally leading innovator in AI, rooted in ethical purpose. Unfortunately, the concept of human-centric may sound as an empty slogan. All commercial activities are virtually human-centric and that doesn't directly guarantee that human and societal factors together to values and rights are either respected or even taken in consideration. For starters, it would be advisable to add 'empowered' to the human, otherwise no human being would be able to the challenge of coping with all the knowledge requirements of facing AI systems. If fake news are a problem for democracy since they weaken the actual awareness of reality, an uninformed person becomes a puppet in front of complex systems that can determine actions on her wealth or her health.
JOSE	VARELA	UGT in collaboration with UNI Europe	- UNI Europa welcomes the possibility to contribute to the stakeholders' consultation and underlines the importance of a broad public debate and information on AI. This debate must result in clear ethical and social guidelines and standards with the aim of improving the living and working conditions of European citizens.- We acknowledge the innovative potential of AI and new technologies that can be beneficial for our society. However, these new technologies also create challenges and we are concerned about the possible risks and consequences relating to working conditions, skills and training, ethics, equality, health and safety (among others). Therefore, UNI Europa would like to underline the importance of addressing AI technologies and robotization as topics for collective bargaining at all levels (company, national and European). AI and robotics have a huge impact on the future labour market, as jobs will sometimes disappear or be transformed and other jobs will be created. We need to accompany this process and address the question of skills and training for the future workforce: need to ensure that training on necessary digital skills is provided by education institutions and companies, and that it is not the sole responsibility of the worker to keep up with	- UNI Europa supports the human-centric or human-in-command approach suggested in the guidelines. We agree that it is necessary that humans always need to remain in control of technology and machines. Likewise, we agree that the use of AI needs to respect European values and fundamental rights.- We recommend an expansion of 3.2 "In an AI context, freedom of the individual requires protection from direct or indirect coercion, surveillance, deception or manipulation." Much profiling that lies at the heart of AI systems relies on a degree of manipulation of data. This is not least relevant in relation to the world of work and especially in the use of AI in Human Resources. - We welcome that the HLEG understands the need to ensure that those involved in the development and marketing of AI (researchers, engineers, designers etc.) act in accordance with ethic and social responsibility criteria. This should be addressed by changing educational priorities for technical subjects and by providing lifelong learning opportunities (e.g. by incorporating ethics and the humanities into training courses in engineering). - Organisations and companies should develop tools to facilitate ethical discussions and decision-making throughout the whole	- In order to achieve "trustworthy AI", we need to establish public, independent and autonomous organisations that can control and audit (labour) algorithms (e.g. to identify underlying biases and the objectivity of data sets that train algorithms). Likewise, the implementation of the ethical guidelines on AI must be monitored. A European observatory focusing on the ethics in AI systems could play the role of an independent watchdog, including in business.- We would like the advice „to always keep record of the data that is fed to the AI systems" from the heading of data governance included under Accountability. For workers, it is paramount that the datasets used to evaluate performance, or in hiring or firing processes is transparent and can be accounted for. - The explanation of the principle of autonomy covers the question of AI at work only in a footnote, whereas this is an important issue that should be given a more prominent place. We would like to highlight the right of workers to individually and collectively opt out or withdraw from the use of AI systems (or a decision chosen by an AI system) if they undermine the workers' autonomy, decision making competence or disrespect fundamental rights and ethical principles. We	- We welcome that processes shall be examined in order "to allow a human control, if needed" (assessment list – governing AI autonomy). In so doing it should not be a question "to keep a human in the loop". We need clearly defined measures, that empower people to exert this control in all processes – regarding resources (technical equipment etc), organisational needs (time, liability, etc.) and qualification.- We suggest extending the list on the assessment of use cases (p.28) and add the question of processes, in order to use AI to ensure decent work (development and impact assessment).	- UNI Europa welcomes the call for Accountability Governance on page 21. The establishment of Data/AI Governance Councils in companies will indeed strengthen the accountability of AI systems and will address a weakness in the GDPR. The Council should consist of shop stewards and management and be responsible for holding management accountable and transparent to the use of AI and data. Whistleblowers should be able to address concerns to the Council and mandate the council to investigate on reported issues. - We welcome the process of developing guidelines for a trustworthy AI made in Europe, which encompasses corresponding "guidelines made in Europe", but would like to raise the question why non-European companies such as Google were granted full membership and full participatory right in the High- Level Expert Group. The status of associate expert would be more appropriate.- UNI Europa also supports the position of the ETUC regarding this consultation.

the rapid technological developments. Employability needs to be promoted through upskilling and reskilling schemes for workers. Investment in formal, informal and life-long learning is key; we must enable people to work with AI or invest in competences that AI will not cover. It is important to develop action plans at EU and national level together with education providers and social partners in order to modernize education and vocational training. We therefore welcome the call from the ILO Global Commission on the Future of Work for "a formal recognition of a universal entitlement to lifelong learning and the establishment of an effective lifelong learning system. (https://www.ilo.org/global/topics/future-of-work/publications/WCMS_662410/lang--en/index.htm)- The social partners play a key role in this and the EU should cooperate with them and national governments in order to identify which job sectors will be affected by AI. We need to understand the timeline and extent of changes in the labour market. The involvement of social partners is a must to find appropriate and future-proof solutions to concerns relating to employment, training, the nature of work, (in)equality or social systems and collective bargaining, especially at sectoral level.- As AI and automation have the potential to transform not only simple tasks but very complex processes, we need to have a large public discussion about the areas in which the use of AI is reasonable and beneficial for society. Part of the debate should be the question of how the profits generated by AI should be re-invested for the common good by creating employment in domains such as care, health services, education or mobility. Employees should participate in the distribution of profits, e.g. through wage increases or reduction of working time. Moreover, AI wins should be used to strengthen social security systems. This could be a measure to address the problem of future job losses and the precarisation of employment relations in a-typical work (e.g. platform work) due to AI and automation.- It is therefore important to integrate the aspect of the quality of jobs, decent work and social progress into the ethical approach in order to create a balance with the purely economic objectives underlying the creation and use of AI and robotics. - The Human-centric approach (HCD) not only presupposes information, transparency, participation and traceability, but also requires specific negotiation processes regarding decision-making in view of the aims and implementation of AI-systems at a very early stage for stakeholders such as employees and their codetermination bodies. design process. This should be completed by internal training programs on ethics for all employees. Such training should help employees understand the AI systems themselves, their rights in relation to said systems and their possibilities of redress, complaint etc. - AI should provide an opportunity for workers to apply their skills and competences to the fullest while at the same time remain owner of the production process. This includes the principle of transparency in the use of AI systems in HR, like the hiring of employees or the performance assessment of staff. It is important to safeguard the rights and freedoms of employees in line with non-discrimination rules as regards the processing of workers' data.- UNI Europa welcomes 5.1 – 5.4. We support that these examples raise real-life concerns of the adverse consequences of AI systems. - In 5.2. UNI Europa urges the group to expand on the issue of the human's right to know they are interacting with an AI identify. This could be done through a "labelling" system. For example, online bots should be labelled as such. Users should be made aware of the use of bots and AI in customer call-centre or help desks etc. - We would welcome that the employer-employee, employer-worker relation is explicitly mentioned in 5.3 as an example of power asymmetry. - Taking into account the power asymmetry in employer-worker relations, a separate point 3.6 on "workers' rights" should be added, which should contain the following points: "decent work by design", equal negotiation processes in the sense codetermination rights, informational self-determination of employees, non-discrimination principle and freedom of association including the right to strike. This is needed in order to secure worker's rights to co-decide on aims and application of AI systems, and create a legal framework.- Concerning the long-term risks and concerns we welcome that these should be considered. This could become an integral part of the accountability and audibility demands – i.e. that developers, users deployers etc need to reflect on the development/changing nature of the adopted AI as well as engage in predictions/forecasts of its future development scope and the consequences (positive and negative) hereof). - AI's influence does not only affect the world of work, but also democracy and society as a whole. We welcome that the draft refers to this point in Chapter I, paragraph 5.3., by stating that AI is not to be implemented in order to enable "citizen scoring" by a state/government. But this should also apply to private businesses. Neither states nor companies should be allowed nor have the possibility, to create human profiles such as "moral personality" or "ethical integrity". We reject the proposed opt-out-function and even possible "opt-in"-functions are not to be designed in a way that they conflict with fundamental human rights and possibly lead to the waiving of services that are useful for a person. AI-based services, that are important for work and life, must be designed in such a way that they do not require the collection of data which could be useable for human profiling.- Creating big data-bases always includes the risk of hackability as well as intentional and unintentional data-leaks. The guiding principle of "data-sovereignty" needs data-security in order to be viable. This recommend the inclusion of a special chapter that provides for ethical guidelines on AI in the work environment to address these issues more in detail.- UNI Europa welcomes that the HLEG on AI acknowledges the importance of social dialogue to realise trustworthy AI. We would like to add that the involvement of social partners, and in particular employee representatives, should not only take place regarding the general public debate on AI. Social partners should be involved in the establishment of codes of conducts, of standardisation schemes, development of training and in the proposed accountability governance. Employee participation and inclusion should take place early in the design, development and deployment of new technologies including AI and robotics. It is essential and important not only to inform and consult workers representatives in the work place or at branch level, but to enhance their co-determination rights and ensure their right to co-decide on the aims, reasons and implementation of AI at the workplace. - Social partners at all levels should be involved in the implementation at company, industry, national or international level, including through collective agreements setting standards. In this context, it is required to describe the negotiation processes, e.g. central control structures for sector-specific solutions (cf. ‚AI Now' Report 2018): „Governments need to regulate AI by expanding the powers of sector-specific agencies to oversee, audit, and monitor these technologies by domain." (https://ainowinstitute.org/AI_Now_2018_Report.pdf)- Regarding the principles of accountability and transparency, we need to establish mechanisms for the protection of whistle-blowers who disclose the risks of AI systems or the non-respect of ethical principles – especially in the case of employees in companies that develop such systems. Internal reporting of risks and violations should be supported and rules in place to ensure follow up. - Organisations and companies should pay attention to potential biases encoded in the system development, training data and model performance – especially those that my affect the most vulnerable. They could also establish an internal ethical review process to democratise the decision-making process- Companies should not only increase transparency regarding the design and development of AI systems, but also in organisational chains of responsibility.

				implies explicitly not surveying data in areas that are of highly explosive nature for people in e.g. political, private or work-related areas. Fundamental rights as informational self-determination, the freedom of association and freedom of speech are not to be put at risk by creating such data-bases.				
Hannah	Wachter	University of Applied Sciences, Kiel	Feedback byProf. Dr. Gaby Lenz (Professor at the University of Applied Sciences Kiel, Faculty of Social Work and Health, scientific accompaniment in the robotics project ARIA, Germany)Hannah Wachter, M.A. (Social worker and project manager at the Austrian Women's Shelter Network, Austria and lecturer at the University of Applied Sciences Kiel, Faculty of Social Work and Health, Germany) General AcknowledgmentsThe following comments are based on a professional understanding of social work, which is based on two pillars: Firstly, the Global Definition of Social Work according to the International Federation of Social Workers, which defines social work as a profession, which promotes social change, development and social justice, based on human rights, empowerment and respect for diversity. The other pillar is the concept of social work as a human rights profession, which is state of the art in the German-speaking discussion (based on the works of the Swiss social work scientist Silvia Staub-Bernasconi). Here, too, the orientation towards human rights, the incorporation of different perspectives (perspectives of social facilities, NGOs, addressees of social work as well as professional standards), political advocacy for vulnerable demographics respectively groups that are affected by biases and a critical view on social developments such as the increasing economization of society.Social work is thus a profession that is confronted and deals with the increasing digitization of society in all its aspects. The impact of the strengthening of AI can be found in many fields, such as domestic violence or labor market policies.The following comments are based on a strong commitment to the European integration process. However, it should also be noted that social integration, such as welfare state standards and the implementation of the Paris Agreement on Climate Change, is currently considered to be subordinate to economic integration and the enlargement of the European single market. From the perspective of social work, in-depth European integration according to social as well as eco-social aspects is strongly recommended.The following comments focus on the following main points in the reception of the AI Ethic Guidelines:• A reinforcement of principles that have already been formulated and that we agree with from a social work based ethical perspective.• A critical discussion of	Chapter I: Respecting Fundamental Rights, Principles and Values – Ethical PurposeThe EU`s Rights` Based Approach on AI EthicsRegarding the orientation of the AI Ethics to the European Convention on Human Rights (see page 5), it should be noted that this orientation is fundamentally welcomed from a social work perspective. It should be noted, however, that this orientation has already begun to falter in recent years due to the (financial) crisis management of the EU, due to the neoliberal austerity policy of recent years. A one-sided focus on austerity policies and lowering social standards is not only questionable in terms of human dignity but can also be considered highly dubious in terms of societal digitization and the current transformation of the labor market. Here a social upheaval is imminent, which urgently requires social state measures to cushion the sometimes disruptive transformation processes.Also, the construction of a „fortress Europe“ with the purpose of foreclosure against refugees is questionable from a social work perspective. It remains a controversial topic in Europe, how to deal with immigration. In our view, the death rates in the Mediterranean Sea as well as the fact that there are still no safe escape routes are to be criticized sharply. Compared to the size of countries like Lebanon and European wealth, the number of refugees could be handled well. However, AI in social media has certainly contributed to the welcome mood 2015 but also its turn into widespread hostility. With regard to the fast-spreading fake news, chatbots-based Hate Spech and algorithm-based information bubbles, it is therefore important to think of solutions to those problems in the HLEG's paper.We note another ethical line of conflict regarding working conditions and environmental pollution in the mining of raw materials, which are necessary for IT and AI. People suffering from precarious work conditions are addressees of social work, and soon climate refugees will increasingly come to Europe. The extraction of raw materials for IT hardware (for example Lithium in Chile or Cobalt in Central Africa) is still highly problematic in terms of working conditions and environmental pollution - a problem that will not be resolved from one day to the next, but which is completely absent in the guidelines so far. An "environmentally friendly AI" (p. 9) is more of a wish in the face of these circumstances than a realizable	Chapter II: Realising Trustworthy AI We support all the requirements of Trustworthy AI. We particularly want to strengthen the focus on data governance, as the implementation of biases and discrimination can be counteracted by a careful selection of the input. Those data sets determine the subsequent logic of the AI.From the Design for all section and the reference to the United Nations Convention on the Rights of Persons with Disabilities, the reference to democratization can also be clarified: "Nothing about us without us." is one of the most important principles of the Convention. As has already been said, it is necessary to emphasize strategies on how people of diverse backgrounds can be included in the transformation process. A particular challenge arises in our view of the participation of people with low social status, since they often have little capacity for participation processes in precarious life situations and are cut off from educational offers. People with disabilities are sometimes exposed to similar problems and exclusion mechanisms too. To tackle those issues, there is a need for social work methodology, e.g. community organizing as well as for the participation of strong interest groups such as the self-advocacy movement.The HLEG has called for more methods to be collected. From our perspective, classical methods such as EU surveys, comment options but also community organizing, open access technology systems as well as the institutionalized involvement of NGOs and civil society organizations are recommended to stimulate stakeholder and social dialogue.At this point we can also support the demands of the draft for education concerning AI (to enable informed consent) as well as for a deeper stakeholder and social dialogues. We also support the demand for diversity and inclusive design teams. An intersectional perspective is recommended to work on hierarchies of power and to prevent a one-sided development of AI. On the one hand, hierarchies of power can be processed between different stakeholder groups, on the other hand we do have power hierarchies within the groups.In regulation / standardization, it is recommended, as already mentioned, to focus on the implementation of the ILO labour standards in the extraction of raw materials for hardware and to tackle the unsolved environmental problems seriously.In the		General CommentsFrom our perspective, the draft on AI Ethics should include an extended perspective. The effects of digitization and the advancement of artificial intelligence on the labor market and on social structures require accompanying social measures. A transformation of labor market policy and a strengthening of the welfare state are urgently needed in view of the impending loss or the transformation of a large number of jobs. Social discussions on paid and unpaid work have to be taken seriously, as well as a questioning of the growth paradigm with regard to environmental challenges (→raw materials, climate change). The ILO labour standards, a discussion on degrowth, a constant reflection on gender issues, and the Paris Agreement can be cornerstones of an extended perspective on AI ethics. The social transformation that has already begun is not to be underestimated in its magnitude, so it needs a comprehensive view and an interdisciplinary, human rights and environmental-oriented discussion that dares real change.

principles that are too vague from our perspective or that do not adequately illuminate the lines of conflict in the development and use of AI.• Suggestions for the creation process of the AI Ethic Guidelines.Introduction: Rationale and Foresight of the GuidelinesAlignment of the guidelinesWe welcome the strong commitment of the HLEG to an AI, which above all pursues a human-centric approach. Above all, the emphasis on an ethical purpose, an orientation towards common good and the dedicated attention to vulnerable demographics and biases is welcomed from the perspective of social work. However, as will be explained below, the paper remains vague in many respects - too vague even if we consider the HLEGs approach to establish general ethical principles and values in the guidelines as a "north star". („tailored approach is needed given AI's context-specificity", p. 3)Strong guidelines should be more specific and obtain a detailed and clear position in the sense of a human-centric approach, in particular concerning trade-offs, for example, between profit-oriented interests and interests of European citizens (→ data collections and data protection). Subsequently, we agree with the HLEG that the legal vacuum, which currently prevails in many AI matters (see page 2), must be cleared up in order to guide European citizens safely and in conformity with human rights into a new digital age. Voluntary guidelines are not enough. Binding legal norms are necessary to reduce the risks of AI to a minimum. At present, laws and norms that govern AI are lagging behind reality in many respects. Examples include gender-based cyber violence such as the use of spyware. While tracking of individuals without consent (rejected by the HLEG on page 11) in issues of domestic violence is increasingly used by violent partners and primarily threatens women, it is largely unclear by law how to handle these violent attacks. At this point, for example, there would have to be a comprehensive ban on the acquisition and use of spyware, as this is highly questionable even for the originally envisaged applications, such as the control of children or employees. Surveillance is an issue that must be used with particular care. Permissible examples would be a baby monitor or moderate monitoring of dementia patients, which leads to more freedom and flexibility. A clear, human rights-compliant regulation is needed here, for example, with the model of a judicially monitored forced placement in psychiatric care because of foreign or self-endangerment. On the other hand, monitoring of adolescents and adults should be categorically rejected, if there`s no need to. Spyware is increasingly being misused for violence in relationships. What is needed here is a rejection of surveillance without consent as a rule (as stated on page 11 of the Draft of the Ethic Guidelines) and, as a result, a clear transfer to binding legal norms and a clear regulation of which products are permitted at all.Review processWe also welcome the fact that fundamentally all addressees who are and will be confronted with AI (individuals, companies, nations and other stakeholders - such as civil society groups) can take the opportunity to influence this review process. Criticism is however noted concerning the period of time: The commenting option was

reality.„Ethical insights help us in understanding how technologies may give rise to different fundamental rights considerations in the development and application of AI, as well as finer grained guidance on what we should do with technology for the common good rather than what we (currently) can do with technology." (p.5)This perspective is explicitly welcomed. From the point of view of social work and considering the currently problematic structures regarding the production of hardware (→ mining of raw materials, raw material conflicts) it must be examined exactly in which fields a deployment makes sense and for which task AI is actually suitable from a human-centric approach perspective. In the words of the expert Meredith Broussard: „It`s time to stop rushing blindly into the digital future and start making better, more thoughtful decisions about when and why to use technology. […] I thought technology was only appropriate if it was the right tool for the right task." (Broussard: Artificial Unintelligence. How computers misunderstand the world 2018, p. 8-9) As an example, the fields of social work and health care can be cited: Humanoid robots can serve complementary to human professionals in very limited fields such as autism or dementia. It could be useful but only if this is not used to further reduce the number of nurses and social workers for cost reasons. In the field of memory performance and the promotion of social skills promising results are achieved in initial research projects. By contrast, the introduction of an algorithm for evaluating labor market opportunities and decisions about trainings for addressees in Austria is, in the unanimous opinion of welfare organizations, diametrically opposed to the social work profession, as the system reflects and reinforces structural forms of discrimination. (see also page 7 - respect for human dignity, rejection of AI, which is bias-based according to the HLEG)From Fundamental rights to Principles and Values„[…] informed consent is a value needed to operationalise the principle of autonomy in practise. Informed consent requires that individuals are given enough information to make an educated decision as to whether or not they will develop, use, or invest in an AI system […]". Basically, we agree with the principle of autonomy and the ability to make our own decisions. In practice, however, there is currently a gap between the requirement for the population to be able to obtain informed consent based on actual knowledge. AI is not a new phenomenon, but the detailed knowledge and the social discourse on AI is still not deep enough, in our opinion, to be able to obtain informed consent. In our opinion, support for individual decision-making processes requires a broad social discourse, participative AI development and decision-making processes with all stakeholders (state, companies, civil society, etc.). The resulting legally anchored norms of protection as an outcome, ultimately must apply to all. Otherwise, an incidence of ethically questionable practices and products is legitimized, with the statement that the users had agreed on the terms. The paper calls for the following: "Organizations should set up an internal or external governance framework to ensure accountability." (P.22) These activities are to be welcomed but

implementation of AI in the area of care and nursing (e.g. care robots) it is essential to install an easy-to-use shut-off button with which robots can be stopped quickly. Furthermore, for reasons of data protection in the sensitive area of care, robots are to be designed that can work offline too.

hardly mentioned in the media and the period is from 18 December 2018 to 1 February 2019 (originally even only until 18 January 2019) conceivable tight. Besides, it`s been also Holiday Season. A serious discussion of the Ethic Guidelines is only partially feasible in this short term. We also consider the participants of HLEG as insufficient. While universities and private-sector companies are predominantly represented, representatives of civil society and NGOs are largely absent. Social structures are therefore poorly modeled and a one-sided representation of interests of private-sector, profit-oriented stakeholders is to be feared. From the point of view of social work, which has a clear objective of (social) participation and empowerment, it is strongly recommended that the assessment process should be more inclusive and participative, also with regard to the current discussion on a democratic deficit of the European Union.

should be subordinate to the norms established in societal participatory processes.Fundamental Rights of Human BeingsThe principles mentioned in this section are central to AI Ethics. In particular, the strong role of democracy in the processes and the reappraisal (and future prevention) of disturbing events in elections in recent years should be given sufficient attention."It should also be noted that, in particular situations, tensions may arise between the principles when considered from the point of view of an individual compared with the point of view of society, and vice versa."In our opinion, there is no clear comment on the tensions and conflicts of values that arise between civic / human rights and profit-oriented interests. Not only individuals and society play a role in these tensions, but also the market and market players. In current classical economics, profit orientation and growth are the maxim. Principles that are not always but proven to often conflict with human rights and environmental protection. Development of AI is currently operated to a large extent by profit-oriented companies, so that profit orientation is also in reverse in the development of AI often in the foreground. Even scientific research institutes are increasingly dependent on third-party funding. Independent financing of development projects and a strengthening of NGOs and other forms of civil society organization are therefore essential in our opinion to prevent a one-sided, profit-oriented orientation of AI and to promote common good.For conflict issues, clear conflict moderation procedures must be established that also adequately address power relations.Covert AI systemsWe welcome that the AI ethics draft also includes the subject covert AI systems. A confusion of humans and AI or the passing of the Turing test exists at present mainly regarding spoken or read products of AI. Although humanoid and zoomorphic robots do not yet look very similar to human beings or animals, they sometimes already induce a similar emotional bond, as Kate Darling explains in her work on Robot Ethics. This aspect - the emotional effect of humanoid and zoomorphic robots - should, in our opinion, be included in the design for AI Ethics and put up for discussion.Already the term "Trustworthy AI" used in this paper implies a confusing anthropomorphizing. Trustworthy AI Design / Development would be more appropriate to increase accountability and make it clear that AI is a human-made system that has no consciousness of its own and only the intelligence of its creators.At this point, apart from the annotations on bias and forms of discrimination, there are also gender issues. From our perspective, rightly, the HLEG places a strong focus on anti-discrimination and reproduced biases in the development and use of AI. The robot is not only considered and used as a machine but conceived as a gendered artifact that reproduces outdated role models in the worst case from a gender-sensitive perspective. Robots that are modeled after unrealistic and clichéd ideals of beauty, the constant availability and utter despotism of assistants with female names like Siri and Alexa (quote: "Let's talk about you.", "What I think is not so interesting, what can I do for you?") reflect social problems and an

				increasing right-wing populist backlash regarding equality policies.		
Jethro	Schiansky	EUnited AISBL				EUnited welcomes the European Commission’s initiative to work on a set of Guidelines concerning the uptake and use of artificial intelligence (AI) in the European Union. It in particular welcomes efforts to define what we mean by the term AI, particularly in the context of ongoing evaluations of existing legislation.Following the publication on the 18th December 2018 of the HLEG on "AI’s Draft Ethics Guidelines for Trustworthy AI" as well as "A Definition of AI: Main Capabilities and Scientific Disciplines", EUnited has the following observations:DefinitionEUnited’s members believe that any definition of AI should describe current and reasonably foreseeable technology and should avoid capturing technology which doesn’t exist and is highly unlikely to exist in the foreseeable future or is commensurate with marketing rather than any available or foreseeable technology. Secondly, EUnited suggests removing all notions from the AI definition which explicitly or implicitly equate machines to humans (phrases and words such as, "perception", “human intelligence”, "reasoning", "interpreting", “reaching conclusions", “learning” etc.). In each case, we offer alternative wording to avoid this risk. As such our proposed definition would be:"Artificial intelligence (AI) refers to computer systems designed by humans that, given a complex task, act by processing the structured or unstructured data collected in their environment according to a set of instructions and operations, determining the best action(s) to take to perform the given task, via software or hardware actuators. AI

								computer systems can also adapt their behaviour by analysing how the environment is affected by their previous actions."Ethics GuidelinesEUnited welcomes many aspects of the Guidelines and the logic of the document described in the introduction is clear and helpful. However, the document remains extremely long and detailed for a set of Guidelines. There may also be some missing concepts in our opinion. Our comments are as follows:• This Guidelines appear to follow the precautionary principle. That is to say there does not appear to be a section dealing with the risks associated with not using AI technology.• Linked to the above, there may also be instances where using AI may be more ethical than not doing so (for example in some medical/healthcare applications), yet this terrain is not explored in the Guidelines.• EUnited understands that Chapter III will be completed by a series of use cases to illustrate "how the framework for trustworthy AI and the Assessment List can be tailored to specific contexts". It is of the utmost importance for an effective ethical Guideline to clearly distinguish areas/applications of high and low risk, rather than simply outlining risks and approaches and treating them in the same way, regardless of the actual risk in the specific field. A one-size-fits-all approach is not suitable here. After all, many AI systems which currently exist operate in applications where there is no contact with humans or impact on them, or they are systems carrying out very simple tasks for which an analysis of all the ethical considerations contained in the Guidelines would be disproportionate. • Linked to the previous point, there are moments where the Guidelines seem overly prescriptive and difficult to imagine being applied in practice. For example, whilst EUnited agrees entirely with the fact that bias is a critical potential problem that must be considered in these Guidelines, prescribing non-discrimination requirements on gender, ethnicity, age, sexual orientation (pg.23) to ensure diversity of teams working on AI systems in companies may be theoretically desirable, but practically unobtainable, particularly for SMEs. The real importance lies in ensuring that AI systems do not lead to discrimination in their application.• In general, it should be borne in mind that AI systems are incredibly broad in terms of the intended application. Ethical considerations should be proportionate to the risks associated with the application, taking into account what rights and obligations are already enshrined in European Law, such as safety and human rights legislation.EUnited remains at the Commission's disposal for more in-depth discussions of the point set out in these comments.
Nicolas	Gomez	Novo Nordisk - Business Assurance				As I see it then transparency in AI is paramount. As the technology advances in such a high-speed guidelines and regulations might struggle to cover all aspects. Therefore, a transparent approach/Open-AI is needed for regulatory bodies to be able to get insight into any potential future implications of a given AI solution. I recommend that this is emphasized as part of the section. It is recommended that the concept of "duty of care" should be emphasized in the	The assessment points in the list is very comprehensive and its apparent that a lot of thought has gone into this. Very well made. Comment to the overall approach: I think an approach to creating a transparent and simple guidance for the assessment is to have an "AI Impact Assessment" which should be completed as a first step. This should serve as a guide for an org. to what detailed controls that are relevant later to have implemented as a mandatory part of their AI solution. The assessment would	I think that the guidelines document is critical component in ensuring that the use of AI in Europe will be able to benefit the individuals life. Further, it will also serve as a guidance for proper security and robustness in AI solutions created which will in turn make it possible to create products that will have competitive edge in the global market. I would like to thank the High level expert group on their excellent work on the guidelines which think overall is a very will written document that has encompassed to

					guidelines. Org. implementing AI have a responsibility to demonstrate duty of care for the individual who is impacted by the solution. This is should be part of a fundamental requirement when creating or operating an AI solution. One example could be AI operated physical robots in a factory should have controls/barriers implemented that physically protect employees working near the robots.	determine ex. based on if the solution has direct impact on people health(digital health solutions) or merely a recommender algorithm(what movies are recommended to watch) which controls to have implemented and to what "control-strength". An assessment would also give internal and regulatory oversight the ability to see what the overall intent is with the AI functionality and if applicable controls have been implemented. Furthermore it will so give the ability to asses if duty of care has been committed by the org. Based on the nature and impact the AI has specific controls and control strength is needed. Based on the AI Impact Assessment the right controls in the control list described in the guideline would be selected. Unfortunately I cannot upload a diagram showing the structure. I would be happy to do so on request. Traceability and audibility: Auditing companies should be able to demonstrate that they are competent to audit AI at a sufficient level since a normal auditing approach is not applicable when dealing with algorithms and data selection and crafting. This could be demonstrated by obtaining ex. certifications.	take a broad use of AI into consideration. If needed I am interested in working in more depth with parts of the document ex. the assessment section since I also have vast experience in assessments, Cyber security and auditing. Thank you!
Katrin	Hatzinger	Protestant Church in Germany (EKD)-Brussels Office	page i, Executive Summary: The fact that the "draft ethics guidelines for trustworthy Artificial Intelligence" (draft) explicitly name opportunities concerning Artificial Intelligence (AI) by concrete examples while remaining unspecific concerning risks, shows a lack of awareness. The draft should be more explicit in naming risks. In its current form the guidelines still have some shortcomings as they seem quite application-oriented and business-friendly.page ii, Executive Summary: The draft guidelines openly talk about "using ethics as an inspiration to develop a unique brand of AI". The wording is ambiguous: If it is understood in the way that ethics is supposed to establish a unique brand of AI this would be equivalent to a utilitarian approach instrumentalizing ethics for predefined goals. If the wording indicates that ethics should explicitly be taken into consideration during the development of AI and should contribute to the trustworthiness of AI (integrated AI-research) it is welcomed. An ethical superiority could pass as an advantage with regard to global competition. The given wording, which stresses the aspect of "competitiveness" in the same context as the "ethical approach to AI" gives reason for our concern that the draft guidelines are driven mainly by business interests and lack a balanced approach. We recommend to clarify the phrasing. page ii, iii, Executive Summary, and page 29, Conclusion: The idea of the draft to establish a kind of European trademark named „Trustworthy AI made in Europe" hinders an open and transparent ethical debate about AI in general. The term and framing of "trustworthy AI" already entail a positive bias towards AI. As a first step a debate about the conditions and procedures how to get trustworthy AI was needed. The draft should put more effort into clarifying that an in-depth reflection is necessary to handle the issue of AI in a responsible way.page iv, glossary "Ethical	page 7, "3. Fundamental Rights of Human Beings: The draft states: „Citizens should enjoy a right to be informed of any automated treatment of their data by government bodies, and systematically be offered to express opt out".Nonetheless, this paragraph does not yet reveal anything about scoring in a private business context. From an ethical perspective it is always problematic to reduce a person's identity to digital data. The guidelines should clarify that it cannot make a difference if governments or business companies abuse artificial intelligence for scoring. The cooperation and links between companies and governments concerning these issues should also be addressed.page 8, Ethical principles in the context of AI and correlating values: We welcome the five mentioned ethical principles and consider them to be of primary importance for an ethical approach to AI. Still there is a need for further clarifications: How do they relate to each other, how should they be balanced against each other when ethical principles are contradicting regarding a technical solution? How does the EU deal with the fact that different actors with different interests might understand ethical principles in a different way? EKD would assume that an ethical discourse based on the rules of fairness and equality needed to be established providing a balanced approach of diverging principles and a common understanding. We would also recommend a human rights impact assemement with regard to algorithms to be established. Surely, public, private, and civil organizations have drawn inspiration from fundamental rights to produce ethical frameworks for AI. The work of European Group on Ethics in Science and New Technologies (EGE) on AI is named in the draft as one example. Nonetheless, the relation between the high-level expert group´s draft guidelines and the EGE opinion on the matter needs to be clarified in the draft. An explanation is needed in which way	page 19, technical methods: The paragraphs on technical methods to realise "trustworthy AI" remain very vague and not specific enough also given the fact that they should apply to "the design, development and use" which is a very broad spectrum. A more differentiated and diligent approach would be needed to look into this important field. page 18, technical and non-technical methods: EKD welcome the fact that the draft guideline stress the importance of an evaluation of the requirements and methods employed on "an on-going basis". page 22, education and awareness to foster an ethical mind-set: Given the fact that EKD considers education on the impact of AI on society as well as on the individuum being of primary importance we feel the draft guidelines should be more specific on this point clarifying that the education should enable the individuum not only to know how to apply AI but provide a broader orientation. page 21, non-technial methods: A human rights impact assessment of AI and the applied algorithms should be added to the list.Page 22, stakeholder and social dialogue: EKD welcomes that the AI HLEG sees the need for "an open discussion and an involvement of social partners, stakeholders and general", but the paragraphs lacks any details about the How of the involvement and remains too vague and unclear with regard to the intended activities. As the involvement of the public is key to establishing a European AI strategy more diligence should be devoted to clarifying this point in EKD´s view also involving theological-ethical expertise.	The present draft constitutes a user-oriented paper addressing developers, deployers and users to comply with fundamental rights and with all applicable regulations. However, it would be necessary to develop a clear concept about the aims to be achieved at the end of the process. It needs to be clarified if there should be a set of regulations with the power to impose sanctions on governments or companies.Ethics is always connected to proceedings and communication channels. Publishing a draft in the pre-Christmas-period remaining open for primarily a period of only one month does not go along with the high-level expert group's declared aim to build ethical guidelines for trustworthy AI in cooperation with the generally public and to allow for a substantive exchange. Moreover, this approach contradicts the ambition of the European Commission to make "the EU more transparent and accountable" through "consultations that are of a high quality and transparent, reach all relevant stakeholders and target the evidence needed to make sound decisions" as stated in the communication "Better regulation for better results – A European Agenda". The circumstances of this opaque procedure give reason for the suspicion that profound debates about this working document are not really desired by the European Commission.We welcome the fact that a voice is given to renowned universities in the European's high-level expert group, but the fact that giant internet companies like google, Zalando or SAP are part of the high-level expert group as well as the biased approach to "a unique brand of AI" without labelling AI as "ethically responsible AI" as well as the content of the guidelines being very application- oriented lead to the impression that the draft is driven by business interests. We also deplore that no theological know-how was involved in the AI HLEG despite competent candidates. Moreover, it should be the aim of such guidelines to avoid the appearance that

purpose": The glossary defines the term "ethical purpose" indicating the "development, deployment and use of AI which ensures compliance with fundamental rights and applicable regulation as well as respecting core principles and values". This paragraph is problematic in several ways. Firstly, because an ethical review starts from a more general perspective and is not purpose-bound in the way that ethics might be (mis-) used as a justification/legitimization for a certain kind of research, development or application. Secondly, the ethical concerns with regard to the development, deployment and use might vary significantly and might even lead to certain contradictions. Therefore, the approach trying to lay out an ethical purpose with regard to "development, deployment and use" must by definition remain quite general and vague and could be misleading. page iv, glossary "Human centric AI": We welcome the human-centric approach to AI. But not only "human values" as mentioned under this heading must be given primary consideration, but human dignity and human rights. This should be clearly mentioned. Page 2, "A. Rationale and foresight of the Guidelines, Purpose and target Audience of the Guidelines": The draft states that „mechanism will be put in place that enable all stakeholders to formally endorse and sign up to the guidelines on a voluntary basis". On the one hand – given the dynamics in AI it is realistic to assume that the guidelines are "a living document" and work in progress, on the other hand there is no clarity about the conclusion of the process. For the time being the guidelines are not legally binding. But what should be the final result of the consultations, ethical reflections etc.? Completely new regulations, an update of the current regulatory framework, a code of conduct, non-binding guidelines? Who are supposed to be the addressees? Who should be held accountable and liable?We want to draw the attention to the risks inherent to an uncoordinated and non-transparent approach in the regulation of AI. A regulatory patchwork may give rise to unclear responsibilities and a lack of accountability leading to a state of bad/ non-governance. The draft should make clear that a ping-pong-effect with regard to accountability and responsibility would not fulfil the requirements of good governance. Moreover, accountability does not have any added value if there is no debate about a regulation of liability at the same time. page 3, "B. A Framework For Trustworthy AI" The draft addresses „developers, deployers and users to comply with fundamental rights and with all applicable regulations". The draft's declared aim is to build guidelines for trustworthy AI. All confidence building processes require as relational acts clear reference objects. The summary and heterogeneous group of „developers, deployers and users" is not appropriate for that purpose. A more tailored, group-specific approach is needed. In addition, it is not clear how the sheer adherence alone to a regulatory framework which is a general obligation for companies and citizens anyway should especially increase the trust in AI compared to the current state of play.

the AI HLEG "build on the above work".page 10, "4. Ethical Principles in the Context of AI and Correlating Values, The Principle of Explicability: The draft explains: „Explicability is a precondition for achieving informed consent from individuals interacting with AI systems and in order to ensure that the principle of explicability and non-maleficence are achieved, the requirement of informed consent should be sought".This statement ignores the fact that the consent in AI touches the principle of human dignity, understood as the recognition of the inherent human state of being worthy of respect. A relational conception of human dignity requires that we are aware of whether and when we are interacting with a machine or another human being for example. We furthermore agree that explicability is a precondition for trustworthy AI. However, the model of „informed consent from individuals" is not realistic for two reasons and therefore should not be applied in this context: on the one hand, the complexity of AI exceeds the capacity of understanding of most individuals. On the other hand, AI algorithms should be legitimately regarded as business secrets which excludes automatically any public explicability. Explicability should be guaranteed towards public authorities and Treuhandstellen (trusts) which are obliged to keep business secrets, dispose of sufficient expertise and which act in the interest of a user or consumer. The present draft does moreover not sufficiently consider the fact, that convergence of interests is not given in the relation between business companies profiting from AI and affected individual human beings. In our view, this aspect of (information) asymmetry must be reconsidered in the draft. Trust in AI will not emerge if AI developers and enterprises refuse to explain their algorithms towards public authorities or state agencies and rely exclusively on their own, interest driven „explications" towards users who usually are lay persons in IT. page 12, Lethal Autonomous Weapon Systems (LAWS): EKD welcomes the fact that regarding lethal autonomous weapon systems (LAWS) the draft clarifies, that "human beings are, and must remain, responsible and accountable for all casualties." However, the draft claims that on the other hand LAWS could "reduce collateral damage, e.g. saving selectively children". Such hypothetical assumptions blur the goal to raise awareness about a responsible handling of AI and seem to justify interests of the military industry which should not be part of the draft.

competition is more important than preservation and – where necessary – enhancement of ethical standards. „Trustworthy AI" can only be successfully established as a brand if the autonomy of the users is strengthened and at the same time an ethically responsible AI is established in correlation with public (governmental or EU) regulation and control in the interest of the users and for the common good . In case this model of AI is refused explicitly by the EU or the enterprises such a concept would even undermine the aim of „trustworthy AI" and hamper the implementation. AI will lead to in-depth societal changes which will go far beyond business-consumer relations and the questions addressed in these guidelines. The risks mentioned under section 5 are not only a question of how to design AI but examples of challenges our societies will have to cope with and find answers to. AI could change and query the current functioning of our societies, the perception of individuals and fundamental principles like human dignity and fundamental rights. Therefore, the EU should set in motion a broad and fundamental societal debate on opportunities and challenges of AI, the relationship of AI to human beings and the society in general.

Sabrina	Zeplin	Otto Group (Otto GmbH & Co. KG)	"Scope of the Guidelines" (p. 3) We appreciate that the present draft explicitly acknowledge that a tailored approach is needed given AI's context-specificity. In regard to higher efficiency and practicability, we would recommend to further sharpen the scope of the AI Ethics Guidelines. We would suggest a differentiation into 'critical' (e.g. Autonomous Driving, Profiling and law enforcement) versus 'non-critical' use cases (e.g. Search Engines, Automation of existing manual Processes). The AI Ethics Guidelines should apply only to use cases that are considered 'critical'. Decisive for whether a use case is considered 'critical' should be which decisions are made based on the algorithm. Accordingly, the potentially made decisions of an algorithm would have to be identified in order to be able to assess their potential damage to humans, animals or the environment. The damage potential could be measured by relatively simple questions, e.g. How many persons are concerned? What is the financial loss? etc. For use cases that were rated 'non-critical' in this context, this should reduce the requested requirements of a Trustworthy AI and their respective assessment.					The Otto Group shares the opinion that no legal vacuum currently exists, as Europe already has regulation in place that applies to AI. However, we very much welcome the AI HLEG´s work on an ethical approach to AI and the draft of European AI Ethics Guidelines that may help limiting regulatory uncertainty in the future. In shaping its digital transformation the Otto Group supports the proposal for a Digital Social Market Economy, which focuses on the freedom and responsibility of individuals, businesses and politics. In our view this should be the European USP that sets us apart from America and Asia."
Jaana	Sinipuro	The Finnish Innovation Fund Sitra	The working document articulates a framework for Trustworthy AI that requires ethical purpose and technical robustness. These are two very important elements, but the most essential element seems to be less emphasized in the document. For AI to be ethical it requires data which must be ethically collected and re-used with permission. Ethically collected data is the raw material for the ethical AI. This feedback concentrates to the personal data and how it can be used within the frame of GDPR. Trustworthy AI isn't born separately in a vacuum. Before data is being enriched by different analytic tools (from which AI is merely one), it first must be collected, organized, modified and stored. All these phases need to be ethically assessed but in an efficient manner which the stakeholders have internalized. The basis for the trustworthiness is built throughout the process of deciding to use AI for any given purpose but it begins from the gathering of the data. There is no functioning AI without the raw material, which the data is. Thus, the emphasis should be placed not only to the use of AI itself but to the whole process beginning with the decision of what kind of data is being gathered, on what premises and authorization. Therefore, ethics should be examined already in the process of gathering the data. Trust is a key factor as the working document very well highlights but to preserve it, the use of AI will require more transparent processes in the future. Mere ethical principles will not be enough. Self-assessment and self-regulation and transparency concerning the use of data and algorithms should be an everyday task for the AI developers, deployers and users.	The Chapter focuses on the core values and principles that all those dealing with AI should comply with. It enhances the principles of protecting individual rights and freedoms while maximizing well-being and the common good. The basic rights of dignity, freedoms, equality and solidary, citizen's rights and justice are the basis as the document brings out. Key Guidance recommends ensuring that AI is human-centric. It should be emphasized that this requires that the used data has been collected on the same basis and principles. Ethical principle of Autonomy and The Question of Ethical Purpose The document mentions Informed consent as a value in the context and use of AI. It is as important in the context of data gathering also. GDPR which came into action in May 2018 requires that the European organizations must ask a permission from the individual when gathering their data. (This however excludes the data which GDPR allows national authorities to collect for specified purposes). Despite the GDPR there are still critical problems which need to be addressed. First, there is the challenge of how the permission is asked. Terms of use are often too wide and difficult for laymen to understand. As the document brings out "consumers give consent without consideration". This document should therefore recommend that the authorization processes should be designed so that the individual can get a good understanding on what the data is gathered and used for and for how long it will be stored. Secondly, unnecessary data should not be gathered at all and the "ethical purpose" should be assessed in this context	Key Guidance for realising Trustworthy AI should be part of the organization culture and part of the everyday life of AI developers, deployers and users. This thinking is well brought out in the chapter. The document brings out that the requirements of Trustworthy AI are all equally important. However, the data governance is an essential starting point for the trustworthy AI. Also, the requirement of transparency has a direct linkage to the data governance requirement. These two should be emphasized since they are the building blocks for the trustworthy and reliable AI. Especially the integrity of the data gathering is important and the decisions on which it is based should be open and transparent. Technical and non-technical Methods to achieve Trustworthy AI Requirements for technical and non-technical methods are comprehensive and manage to consider the variety of procedures the trustworthy AI requires. As it is acknowledged in the document, AI needs to be secure with its processes, data and outcomes and to take adversial data and attacks into account.	-	Document offers solid guidelines and a good base for the use of trustworthy AI. We have evaluated the document from the perspective of fair use of data within the context of GDPR, which we consider to be the starting point for the trustworthy AI. The document considers the whole process and emphasizes the on-going nature of evaluation for trustworthy AI, but it seems to pay less attention to the gathering of the data. We want to bring out that there should be a stronger linkage between the fair data use and the trustworthy AI. The Finnish Innovation Fund's project Human-Driven Data Economy aims to build the foundation for a fair and functioning data economy. The main objectives are to create a method for data exchange and to set up European level rules and guidelines for ethical use of data.

			The new mechanism should therefore emphasize the whole process and take account the data gathering as a critical phase in creating / using trustworthy AI.	also. The question of ethical purpose is directly linked to the question of the amount of data which is being gathered all the time. It should be recognized that the on-going collection of data from our everyday lives is starting to have effect on our right to the privacy and anonymity. We might even find that people are starting to limit their actions just because they want to restrict the data collected from them. This may have direct impacts to individual freedoms and the society at large. One question, that should be considered in societal decision-making in particular, relates to the sampling of data collection. How can it be ensured that the data used in analyses is not already biased? Groups that use digital services less than average, such as those in the weakest position in society, might be excluded from data collection. Unless attention is paid to this, social decisions based on the analysis of data might lead to an even greater deterioration in the position of the individuals outside the scope of data collection.These kinds of questions are important to consider in the context of AI. The principle of ethical purpose supports the fundamental rights of human beings (respect for human dignity, freedom of the individual, respect for democracy, justice and the rule of law, equality, non-discrimination and solidarity and citizen rights).		
Oliver	Suchy	Deutscher Gewerkschaftsbund (DGB)	Es ist zu begrüßen, dass der Entwurf der HLEG dem Ansatz folgt, KI mit europäischen Werten und Prinzipien zu verbinden. Dabei ist besonders hervorzuheben, dass ethische Grundsätze nicht nur benannt werden, sondern auch in konkrete Handlungsempfehlungen / Leitlinien bei der Entwicklung und der Umsetzung von KI münden.Die formulierten Zielvorstellungen – „do good, do no harm, autonomy of humans, justice, and explicability" – sind jedoch sehr allgemein, idealtypisch und im Hinblick auf die Arbeitswelt unzureichend. Bei der Zielsetzung im Sinne von „human centric" wird dem Einsatz von KI ein ethischer Zweck („ethical purpose") zugeordnet, der neben der technischen Zuverlässigkeit von KI-Systemen („technically robust and reliable") als tragende Säule einer „vertrauenswürdigen KI made in Europe" gelten soll.Grundsätzlich ist zu bedenken, dass KI von Menschen gemacht wird – und damit immer unterschiedlichen Interessen folgt. Weitgehend ausgeblendet werden aber – auch im weiteren Verlauf des Entwurfs der Ethik-Richtlinien – ökonomische und strategische Fragen zum Einsatz von KI in Bezug auf die Veränderung von Wertschöpfung, Wirtschaftsstrukturen, Arbeitsmärkten und Beschäftigung) oder auch arbeitspolitischen Grundsätzen (z. B. individuelle Optimierung und Verantwortung, insbesondere im Arbeits- und Gesundheitsschutz). KI kann zweifellos dazu beitragen, gesellschaftliche Herausforderungen (Gesundheit, Umwelt oder Klima) zu bewältigen. Jedoch ist auf europäischer Ebene nicht davon auszugehen, dass wirtschaftliche Interessen zur Anwendung und Verwertung von KI-Systemen generell deckungsgleich mit	Entscheidend für die Wirkungsweise von KI-Systemen ist die Frage der Entwicklung von Zielen, denen KI-Systeme folgen. Dies wird im Entwurf der HLEG nicht thematisiert und sollte dringend ergänzt werden.Es wird zwar darauf hingewiesen, grundlegende Prinzipien zu beachten (Accountability, Data Governance, Design for all, Governance of AI Autonomy (Human oversight), Non-Discrimination, Respect for Human Autonomy, Respect for Privacy, Robustness, Safety, Transparency). Es wird auch darauf hingewiesen, Stakeholder bei der Konzeption und Entwicklung von KI zu beteiligen, wobei aber anzumerken ist, dass alle Stakeholder inkl. insbesondere der Beschäftigten beteiligt werden müssen („Ensure participation and inclusion of stakeholders in the design and development of the AI system") – analog zu Kap. II bei der Rolle von Bildung und Sensibilisierung bei der Sicherstellung von vertrauenswürdiger KI (wo es heißt: „Trustworthy AI requires informed participation of all stakeholders").Es ist zu begrüßen ist, dass der soziale Dialog als wichtiges Element benannt wird, denn der wirtschaftliche Erfolg hängt sehr eng mit der gesellschaftlichen Akzeptanz, nicht zuletzt bei den Beschäftigten zusammen. Hier ist jedoch eine Beschreibung der Verhandlungsprozesse erforderlich, wie zum Beispiel zentrale Kontrollstrukturen für Branchenlösungen (vgl. ‚AI-Now'-Bericht 2018): „Governments need to regulate AI by expanding the powers of sector-specific agencies to oversee, audit, and monitor these technologies by domain." (https://ainowinstitute.org/AI_Now_2018_Report.pdf). Problematisch ist jedoch, dass sich der Beteiligungsansatz nicht explizit auf die Entwicklung der Ziele des KI-Systems	Wie bereits in den Anmerkungen zu Chapter I beschrieben, mangelt es an der Beschreibung von Verhandlungsprozessen, die zur Erreichung der Zielsetzung eines „human centered design" bedürfen. KI wird von Menschen gemacht und folgt somit auch unterschiedlichen Interessen.Der Ansatz, die o. g. Leitlinien bereits am Beginn der Entwicklung von KI-Systemen zu berücksichtigen und den Einsatz von KI als kontinuierlichen Prozess ist zu begrüßen. Die Beschränkung auf „Information und Beteiligung der Stakeholder" ist jedoch unzureichend. Erforderlich sind partnerschaftliche Aushandlungsprozesse sowie eine Dokumentation der Optimierungsziele und deren kontinuierliche Evaluation. Dafür sind die Mitbestimmungsrechte zu verbessern, denn es geht nicht nur um das Einholen von Meinungen, sondern es geht darum, bei dem „was, wozu und wie" des Einsatzes von KI mitzuentscheiden.Die Sicherstellung der Nachvollziehbarkeit (Traceability) von der Wirkungsweise von KI-Anwendungen ist zu begrüßen, erfordert jedoch auch Interventionsmöglichkeiten. Der Vorschlag für ein Opt-Out (nach dem Entwurf im Bereich Scoring) ist für Beschäftigte im betrieblichen Kontext keine realistische Option. Vielmehr sind die Überwachungs-, Optimierungs- und Vorhersagemöglichkeiten menschlicher Arbeit bzw. Leistungsfähigkeit durch KI zu beachten und negative Konsequenzen für Beschäftigte gemäß Art. 22 DSGVO und Erwägungsgrund 71 DSGVO auszuschließen. Das Letztentscheidungsrecht muss immer beim Menschen liegen. Klärungsbedürftig ist auch die Frage der Verantwortung im Umgang mit Entscheidungsvorschlägen von KI-Systemen Es ist zu begrüßen, dass konkrete Ansätze und Fragestellungen aufgelistet werden, um KI-Anwendungen kontinuierlich zu monitoren und zu bewerten.An der – offenen – Frage zur Verantwortung und zum Umgang mit potenziellen Problemen oder Risiken zeigt sich exemplarisch die Notwendigkeit, gemeinsam verbindliche Prozesse der Überprüfung zu vereinbaren (s. Anmerkungen zu Chapter II).So ist es zwar zu begrüßen, dass Prozesse darauf geprüft werden sollen „to allow a human control, if needed" (Assessment List Punkt 4. „Governing AI autonomy"). Dabei darf es aber nicht darum gehen „to keep a human in the loop". Es bedarf klarer Maßnahmen, die den Menschen sowohl hinsichtlich der Ressourcen (technische Vorrichtungen u.ä.) und organisatorisch (Zeit, Haftung, etc.) als auch qualifikatorisch in allen Prozessen befähigen, diese Kontrolle auszuüben.Es wird vorgeschlagen, die geplante Liste zur Bewertung der use cases (p. 28) um die Frage nach den Prozessen zu erweitern, um KI für Gute Arbeit zu nutzen (Entwicklung und Folgenabschätzung).	Es ist zu begrüßen, dass Regulierungsfragen in einem weiteren Schritt diskutiert werden sollen. Um das Ziel zu erreichen, „ein Klima für Innovation und Akzeptanz von KI zu fördern", sollte der rechtliche Rahmen für Aushandlungsprozesse und Gestaltungsoptionen von KI in der Arbeitswelt auf die Agenda gesetzt werden. Die Akzeptanz bei den Beschäftigten ist eine entscheidende Sollbruchstelle für die Umsetzungsaussichten von KI in Unternehmen.Zu begrüßen ist auch, dass kritische Anmerkungen formuliert worden sind, wenngleich dazu kein Konsens innerhalb der HLEG gefunden werden konnte. Um Risiken zu minimieren und damit die Chancen für gesellschaftlichen Fortschritt zu erhöhen, ist eine offene Diskussion erforderlich. Dies betrifft das Verhältnis von Mensch und KI – und letztlich immer die Frage, wofür KI eingesetzt wird und wer über die Optimierungsziele entscheidet. In dem Zusammenhang stellt sich jedoch die Frage, weshalb bei der Entwicklung von Leitlinien, die ausdrücklich zu einer „vertrauenswürdigen KI Made in Europe" – also auch entsprechende Leitlinien Made in Europe – führen sollen, auch außereuropäische Unternehmen wie z.B. Google nicht lediglich als Sachverständige angehört wurden, sondern eine direkte Mitgliedschaft mit entsprechendem Mitspracherecht in der HLEG erhielten. • Grundsätzliche Anmerkung/FundamentalsGrundsätzlich wirkt sich der Einfluss von KI sowohl auf Arbeitswelt als auch auf Demokratie und Gesellschaft aus. Es ist zu begrüßen, dass in Kap. I Punkt 5.3 Bezug darauf genommen wird, dass der Einsatz Künstlicher Intelligenz keinem großangelegtem staatlichen „citizen

politischen Ansprüchen zur Lösung gesellschaftlicher Probleme sind.Deshalb ist im Hinblick auf den Einsatz von KI-Systemen in der Arbeitswelt eine ethische Komponente zur Förderung von Guter Arbeit und sozialem Fortschritt zu ergänzen. Dieser Grundsatz gilt insbesondere im öffentlichen Bereich, der Vorbildfunktion für die kommerzielle Anwendung bzw. wirtschaftliche Nutzung von KI haben sollte. Dies gilt auch und nicht zuletzt für die notwendigen Prozesse zur Entwicklung und Umsetzung von „trustworthy AI". „Gute Arbeit by design" sollte als Grundprinzip für den Einsatz von KI in der Arbeitswelt gelten. Dies ist eine logische Konsequenz aus der Zielsetzung der HLEG, nach der KI kein Selbstzweck ist, sondern das menschliche Wohlbefinden verbessern soll. Dies sollte nicht zuletzt für das Arbeitsleben gelten.Der „human centric"-Ansatz (HCD) setzt nicht nur Information, Transparenz, Nachvollziehbarkeit und Beteiligung voraus, sondern erfordert kontextspezifische Aushandlungsprozesse im Sinne des frühzeitigen Mit-Entscheidens von Stakeholdern wie der Beschäftigten und ihrer Interessenvertretungen über Ziele und Umsetzung von KI-Systemen.

bezieht, obwohl konstatiert wird, dass Zielkonflikte („fundamental tensions between different objectives") bestehen können.Es ist unzureichend, diese Trade-offs nur zu kommunizieren und zu dokumentieren, wie es im Entwurf vorgeschlagen wird. Vielmehr ist es für die Umsetzung und Akzeptanz von entscheidender Bedeutung, dass Optimierungsziele und mögliche Zielkonflikte bereits vor der Implementation von KI-Systemen verhandelt und gelöst werden. Dies gilt insbesondere für den Einsatz von KI in der Arbeitswelt.Die HLEG beschreibt im Entwurf zwar die Notwendigkeit, Machtverhältnissen besondere Beachtung zu schenken („asymmetries of power or information") und nennt hier exemplarisch das Verhältnis von Arbeitgebern und Beschäftigten.Gleichwohl wird (in Chapter II) kein Prozess empfohlen, der auf gleichberechtigten Aushandlungsprozessen der Stakeholder beruht. Die Verhandlung von Optimierungszielen für ein KI-System im betrieblichen Kontext ist jedoch eine zwingende Voraussetzung, um (a) das Erfahrungswissen der Beschäftigten für die Prozessoptimierung zu nutzen, (b) die nutzer- bzw. arbeitnehmerfreundliche Umsetzung zu erleichtern, mögliche Zielkonflikte zu lösen, (d) die Akzeptanz bei den Beschäftigten zu erhöhen und damit (e) Synergien für eine erfolgreiche Umsetzung im Sinne von Guter Arbeit (Verbesserung der Arbeitsbedingungen) und wirtschaftlichem Erfolg (Erhöhung der Produktivität) – oder auch gesellschaftlichem Nutzen – zu erreichen.Die Aushandlungsprozesse sind von besonderer Bedeutung, wenn KI-Systeme auf persönlichen Daten von Beschäftigten basieren oder diese tangieren. Information und Beteiligung sind hier ebenso unzureichend wie das Prinzip der informierten Einwilligung („informed consent"), das angesichts der Machtasymmetrie von abhängig Beschäftigten ohnehin kein Maßstab im Arbeitsleben sein kann.Um den Machtasymmetrien in Beschäftigungs- und Auftragsverhältnissen insgesamt Rechnung zu tragen, sollte daher unter den Fundamental Rights of Human Beings ein eigener Punkt 3.6 zu „Workers rights" aufgenommen werden. Darunter sind insbesondere „Gute Arbeit by design", gleichberechtigte Aushandlungsprozesse im Sinne von Mitbestimmungsrechten, informationelle Selbstbestimmung von Beschäftigten (bzw. Erwerbstätigen in Abhängigkeitsverhältnissen), Diskriminierungsverbot, sowie insbesondere das Recht auf Vereinigungsfreiheit inkl. des Rechts auf Streik aufzuführen, um effektive Machtressourcen Beschäftigter beim Mit-Entscheiden über Ziele und Einsatzweisen von KI rechtlich-institutionell abzusichern.

für Beschäftigte.Es ist zu begrüßen, dass die HLEG die Bedeutung von Aus- und Weiterbildung anerkennt. Insbesondere die Ausbildung von KI-Entwicklern sollte die ethischen Prinzipien für KI beinhalten. Zu kritisieren ist allerdings, dass Beschäftigte (außer im Kontext der Anwendung (users) nicht berücksichtigt sind. Darüber hinaus sollte es insbesondere im Arbeitskontext nicht nur um die Ausbildung für KI gehen, sondern um die Qualifizierung für den Einsatz und Umgang mit KI-Systemen (Veränderung von Tätigkeitsprofilen etc.).Dafür ist eine technische und soziale Folgenabschätzung von KI-Systemen am Beginn der Entwicklung und in der Umsetzung erforderlich, die sich neben Qualifizierungserfordernissen auch auf Belastungsveränderungen und die Frage der Datennutzung bezieht. Menschenzentrierte KI bedeutet auch, dass eine Richtlinie Stellung dazu bezieht, unter welchen Bedingungen z. B. Trainingsdatensätze generiert werden.

scoring" dienen darf. Allerdings sollte dies auch in Bezug auf private Unternehmen gelten. Sowohl Staaten als auch Unternehmen sollte es weder erlaubt noch möglich sein, menschliche Profilbildungen wie bspw. „moral personality" oder „ethical integrity" zu erstellen. Die hier vorgeschlagenen „Opt-out"-Funktionen sind klar abzulehnen und selbst mögliche „Opt-In"-Funktionen dürfen nicht so gestaltet sein, dass sie Grundrechten widersprechen und mit einem Verzicht auf für den Menschen nützliche Dienstleistungen einhergehen. Für Arbeit und Leben wichtige und nützliche KI-basierte Dienstleistungen müssen so gestaltet werden, dass sie keine Datenerhebungen voraussetzen, die für menschliche Profilbildungen nutzbar wären.Die Erstellung großer Datensätze ist immer mit dem Risiko der Hackbarkeit (hackability) und beabsichtigter wie unbeabsichtigter Leaks verbunden. Insofern ist auch die Leitidee von „Datensouveränität" nur tragfähig, wenn dieser Sicherheitsaspekt mitgedacht wird. Das bedeutet, strukturelle Maßnahmen der Dezentralisierung bzw. in manchen – für den Menschen politisch, privat oder arbeitstechnisch brisanten – Bereichen auch die ausdrückliche Nichterhebung von Daten festzulegen. Das Grundrecht auf informationelle Selbstbestimmung, auf Meinungs- und Koalitionsfreiheit darf nicht durch die Schaffung solcher Datenbasen gefährdet werden.

Anonymous Anonymous Anonymous

These guidelines are a very welcome initiative for the EU to embrace the new general-purpose technology Artificial Intelligence and to proactively provide guidance on the development and deployment of such technology. I endorse the guidelines on Trustworthy AI in full. However, I want to draw attention to one mission dimension: it is not only crucial WHAT technology is developed but also WHO gets to use it. Thus, I make a proposal to add a DUE DILLIGENCE process to the non-technical methods for trustworthy AI.

Even if a company aims to develop technology with an ethical purpose and abides by the technical and non-technical methods for trustworthy AI, it might have a blind spot: the DUAL USE potential of AI technology will always remain. This means that while the company may intend clients to only use it for benevolent applications, malicious actors could also use it to inflict harm onto others. Thus, companies have a DUE DILLIGENCE responsibility in deciding WHOM to sell their technology.

Let me illustrate this using an example problem from a Dutch tech company in which I used to work. A facial recognition software product (mostly used by universities for research) was sold globally via a reseller and when reviewing the client list at some point it was noticed that the Chinese Ministry of Public Security had bought the facial recognition software as well. This ministry has a track record of systematic human rights violatons and has a declared goal to create an “omnipresent, completely connected, always on and fully controllable” national video surveillance network. While the company did not intend to provide technology for the purposes of the ministry, its AI technology nevertheless ended up in the hands of an authoritarian surveillance state.

In hindsight, this technology was sold due to a lack of DUE DILLIGENCE, defined as the investigation or exercise of care that a reasonable business or person is expected to take before entering into an agreement or contract with another party, or an act with a certain standard of care. Since AI is a powerful technology this responsibility to evaluate in context who could use it maliciously is very high. Publications, such as ‘The Malicious Use of Artificial Intelligence - Forecasting, Prevention, and Mitigation’ (2018, by Brundage et al.), make a strong case that it matters not only WHAT and HOW forms of AI technology are DEVELOPED, but also WHO gets to BUY and USE them in which CONTEXT. While it may be admissible to use a facial recognition software for experiments approved by the ethics board of a university in the EU, it may not be acceptable to use THE SAME TECHNOLOGY in the context of the Chinese surveillance state. While at times this is hinted at in the current draft guidelines, it is nowhere made explicit and the resulting DUE DILLIGENCE responsibility is missing on the lists for methods for trustworthy AI.

Thus, I strongly urge the members of the High-Level Expert Group on AI to add DUE DILLIGENCE processes to the non-technical methods for AI. This could include but is not limited to: a) creating awareness and

					processes in companies to consider the potential dual use and malicious applications of their technology b) adding clauses in sales and reselling agreements limiting the use of AI to legal applications with an ethical purpose and specifically excluding prominent malicious applications, c) background checks performed by sales personnel when selling sensitive products, d) excluding certain actors ex ante from becoming potential customers, for instance those governments who have been found to systematically violate human rights (this list is to be specified and updated based on academic sources and authoritative indices). In conclusion, I urge the members of the High-Level Expert Group to include DUE DILLIGENCE as a non-technical method for trustworthy AI, to ensure that powerful technology does not land in the hand of those aiming to use it in contradiction to the shared values of the European Union.		
Chevet	Stéphane	CFDT F3C	- UNI Europa ICTS and Union CFDT F3C welcomes the possibility to contribute to the stakeholders' consultation and underlines the importance of a broad public debate and information on AI. This debate must result in clear ethical and social guidelines and standards with the aim of improving the living and working conditions of European citizens.- We acknowledge the innovative potential of AI and new technologies that can be beneficial for our society. However, these new technologies also create challenges and we are concerned about the possible risks and consequences relating to working conditions, skills and training, ethics, equality, health and safety (among others). Therefore, UNI Europa would like to underline the importance of addressing AI technologies and robotization as topics for collective bargaining at all levels (company, national and European). AI and robotics have a huge impact on the future labour market, as jobs will sometimes disappear or be transformed and other jobs will be created. We need to accompany this process and address the question of skills and training for the future workforce: need to ensure that training on necessary digital skills is provided by education institutions and companies, and that it is not the sole responsibility of the worker to keep up with the rapid technological developments. Employability needs to be promoted through upskilling and reskilling schemes for workers. Investment in formal, informal and life-long learning is key; we must enable people to work with AI or invest in competences that AI will not cover. It is important to develop action plans at EU and national level together with education providers and social partners in order to modernize education and vocational training. We therefore welcome the call from the ILO Global Commission on the Future of Work for "a formal recognition of a universal entitlement to lifelong learning and the establishment of an effective lifelong learning system. (https://www.ilo.org/global/topics/future-of-work/publications/WCMS_662410/lang--en/index.htm)- The social partners play a key role in this and the EU should cooperate with them and national governments in order to identify which job sectors will be	- UNI Europa supports the human-centric or human-in-command approach suggested in the guidelines. We agree that it is necessary that humans always need to remain in control of technology and machines. Likewise, we agree that the use of AI needs to respect European values and fundamental rights.- We recommend an expansion of 3.2 "In an AI context, freedom of the individual requires protection from direct or indirect coercion, surveillance, deception or manipulation." Much profiling that lies at the heart of AI systems relies on a degree of manipulation of data. This is not least relevant in relation to the world of work and especially in the use of AI in Human Resources. - We welcome that the HLEG understands the need to ensure that those involved in the development and marketing of AI (researchers, engineers, designers etc.) act in accordance with ethic and social responsibility criteria. This should be addressed by changing educational priorities for technical subjects and by providing lifelong learning opportunities (e.g. by incorporating ethics and the humanities into training courses in engineering). - Organisations and companies should develop tools to facilitate ethical discussions and decision-making throughout the whole design process. This should be completed by internal training programs on ethics for all employees. Such training should help employees understand the AI systems themselves, their rights in relation to said systems and their possibilities of redress, complaint etc. - AI should provide an opportunity for workers to apply their skills and competences to the fullest while at the same time remain owner of the production process. This includes the principle of transparency in the use of AI systems in HR, like the hiring of employees or the performance assessment of staff. It is important to safeguard the rights and freedoms of employees in line with non-discrimination rules as regards the processing of workers' data.- UNI Europa welcomes 5.1 – 5.4. We support that these examples raise real-life concerns of the adverse consequences of AI systems. - In 5.2. UNI Europa urges the group to expand on the issue of the human's right to know they are interacting with an AI identify. This	- In order to achieve "trustworthy AI", we need to establish public, independent and autonomous organisations that can control and audit (labour) algorithms (e.g. to identify underlying biases and the objectivity of data sets that train algorithms). Likewise, the implementation of the ethical guidelines on AI must be monitored. A European observatory focusing on the ethics in AI systems could play the role of an independent watchdog, including in business.- We would like the advice „to always keep record of the data that is fed to the AI systems" from the heading of data governance included under Accountability. For workers, it is paramount that the datasets used to evaluate performance, or in hiring or firing processes is transparent and can be accounted for. - The explanation of the principle of autonomy covers the question of AI at work only in a footnote, whereas this is an important issue that should be given a more prominent place. We would like to highlight the right of workers to individually and collectively opt out or withdraw from the use of AI systems (or a decision chosen by an AI system) if they undermine the workers' autonomy, decision making competence or disrespect fundamental rights and ethical principles. We recommend the inclusion of a special chapter that provides for ethical guidelines on AI in the work environment to address these issues more in detail.- UNI Europa ICTS welcomes that the HLEG on AI acknowledges the importance of social dialogue to realise trustworthy AI. We would like to add that the involvement of social partners, and in particular employee representatives, should not only take place regarding the general public debate on AI. Social partners should be involved in the establishment of codes of conducts, of standardisation schemes, development of training and in the proposed accountability governance. Employee participation and inclusion should take place early in the design, development and deployment of new technologies including AI and robotics. It is essential and important not only to inform and consult workers representatives in the work place or at branch level, but to enhance their co-determination rights and ensure their right to co-decide on the aims, reasons and	- We welcome that processes shall be examined in order "to allow a human control, if needed" (assessment list – governing AI autonomy). In so doing it should not be a question "to keep a human in the loop". We need clearly defined measures, that empower people to exert this control in all processes – regarding resources (technical equipment etc), organisational needs (time, liability, etc.) and qualification.- We suggest extending the list on the assessment of use cases (p.28) and add the question of processes, in order to use AI to ensure decent work (development and impact assessment).	- UNI Europa ICTS welcomes the call for Accountability Governance on page 21. The establishment of Data/AI Governance Councils in companies will indeed strengthen the accountability of AI systems and will address a weakness in the GDPR. The Council should consist of shop stewards and management and be responsible for holding management accountable and transparent to the use of AI and data. Whistleblowers should be able to address concerns to the Council and mandate the council to investigate on reported issues. - We welcome the process of developing guidelines for a trustworthy AI made in Europe, which encompasses corresponding "guidelines made in Europe", but would like to raise the question why non-European companies such as Google were granted full membership and full participatory right in the High- Level Expert Group. The status of associate expert would be more appropriate.- UNI Europa also supports the position of the ETUC regarding this consultation.

affected by AI. We need to understand the timeline and extent of changes in the labour market. The involvement of social partners is a must to find appropriate and future-proof solutions to concerns relating to employment, training, the nature of work, (in)equality or social systems and collective bargaining, especially at sectoral level.- As AI and automation have the potential to transform not only simple tasks but very complex processes, we need to have a large public discussion about the areas in which the use of AI is reasonable and beneficial for society. Part of the debate should be the question of how the profits generated by AI should be re-invested for the common good by creating employment in domains such as care, health services, education or mobility. Employees should participate in the distribution of profits, e.g. through wage increases or reduction of working time. Moreover, AI wins should be used to strengthen social security systems. This could be a measure to address the problem of future job losses and the precarisation of employment relations in a-typical work (e.g. platform work) due to AI and automation.- It is therefore important to integrate the aspect of the quality of jobs, decent work and social progress into the ethical approach in order to create a balance with the purely economic objectives underlying the creation and use of AI and robotics. - The Human-centric approach (HCD) not only presupposes information, transparency, participation and traceability, but also requires specific negotiation processes regarding decision-making in view of the aims and implementation of AI-systems at a very early stage for stakeholders such as employees and their codetermination bodies.

could be done through a "labelling" system. For example, online bots should be labelled as such. Users should be made aware of the use of bots and AI in customer call-centre or help desks etc. - We would welcome that the employer-employee, employer-worker relation is explicitly mentioned in 5.3 as an example of power asymmetry. - Taking into account the power asymmetry in employer-worker relations, a separate point 3.6 on "workers' rights" should be added, which should contain the following points: "decent work by design", equal negotiation processes in the sense codetermination rights, informational self-determination of employees, non-discrimination principle and freedom of association including the right to strike. This is needed in order to secure worker's rights to co-decide on aims and application of AI systems, and create a legal framework.- Concerning the long-term risks and concerns we welcome that these should be considered. This could become an integral part of the accountability and audibility demands – i.e. that developers, users deployers etc need to reflect on the development/changing nature of the adopted AI as well as engage in predictions/forecasts of its future development scope and the consequences (positive and negative) hereof). - AI's influence does not only affect the world of work, but also democracy and society as a whole. We welcome that the draft refers to this point in Chapter I, paragraph 5.3., by stating that AI is not to be implemented in order to enable "citizen scoring" by a state/government. But this should also apply to private businesses. Neither states nor companies should be allowed nor have the possibility, to create human profiles such as "moral personality" or "ethical integrity". We reject the proposed opt-out-function and even possible "opt-in"-functions are not to be designed in a way that they conflict with fundamental human rights and possibly lead to the waiving of services that are useful for a person. AI-based services, that are important for work and life, must be designed in such a way that they do not require the collection of data which could be useable for human profiling.- Creating big data-bases always includes the risk of hackability as well as intentional and unintentional data-leaks. The guiding principle of "data-sovereignty" needs data-security in order to be viable. This implies explicitly not surveying data in areas that are of highly explosive nature for people in e.g. political, private or work-related areas. Fundamental rights as informational self-determination, the freedom of association and freedom of speech are not to be put at risk by creating such data-bases.

implementation of AI at the workplace. - Social partners at all levels should be involved in the implementation at company, industry, national or international level, including through collective agreements setting standards. In this context, it is required to describe the negotiation processes, e.g. central control structures for sector-specific solutions (cf. ‚AI Now' Report 2018): „Governments need to regulate AI by expanding the powers of sector-specific agencies to oversee, audit, and monitor these technologies by domain." (https://ainowinstitute.org/AI_Now_2018_Report.pdf)- Regarding the principles of accountability and transparency, we need to establish mechanisms for the protection of whistle-blowers who disclose the risks of AI systems or the non-respect of ethical principles – especially in the case of employees in companies that develop such systems. Internal reporting of risks and violations should be supported and rules in place to ensure follow up. - Organisations and companies should pay attention to potential biases encoded in the system development, training data and model performance – especially those that my affect the most vulnerable. They could also establish an internal ethical review process to democratise the decision-making process- Companies should not only increase transparency regarding the design and development of AI systems, but also in organisational chains of responsibility.

Shameek	Kundu	Standard Chartered Bank	Broadly, we are supportive of the Guidelines. We appreciate that the Artificial Intelligence High Level Expert Group (AI HLEG) has taken inspiration from similar initiatives in other jurisdictions (such as in the UK and Canada in particular) when developing the Guidelines, as we strongly believe in the benefits of international convergence. Moreover, if sector-specific approaches are to be developed using these Guidelines, for the financial sector we would recommend seeking alignment with the Monetary Authority of Singapore's Principles to promote fairness, ethics, accountability and transparency (FEAT) in the use of artificial intelligence and data analytics in the financial sector. We would also recommend that any sector-specific approaches should be undertaken by the relevant regulator for that sector. Such consistency will be key for financial institutions operating on a global basis. The AI HLEG acknowledges there are tensions and necessary trade-offs between the non-exhaustive list of the 10 requirements of trustworthy AI. Moreover, the AI HLEG recognises that the use of AI raises different challenges in different situations and industries, and therefore explicitly acknowledges that 'a tailored approach is needed given AI's context-specificity'. It is not immediately clear, however, how this tailoring will be taken forward or applied beyond the four specific use-cases listed on page 28, to be developed in the final version of the Guidelines. We recommend that the AI HLEG provide guidance with specific case studies illustrating how tensions and trade-offs between the requirements are dealt with in practice. We would welcome clarification on the Policy & Investment Recommendations, particularly with respect to its scope, approach and intended legal or regulatory form, especially as it is mentioned in the Guidelines that "no legal vacuum currently exists, as Europe already has regulation in place that applies to AI". We would encourage any future work to take a pragmatic, flexible and principles-based approach to ensure that the overall European framework for trustworthy AI remains futureproof and does not unduly constrain innovation. It is mentioned that a final version of the Guidelines will put forward a mechanism for voluntary endorsement by stakeholders. We would appreciate further details on the process and mechanism for this, as well as on: whether further consultation on the mechanism will be held; whether or how stakeholders will be held to account on their endorsement; or if partial endorsement will be possible, which may be required recognising that "there might be fundamental tensions between different objectives".	With regard to the principle of Autonomy: "Preserve Human Agency", the Guidelines refer to consumers or users of an AI system having the right to decide whether they wish to be subject to AI decision-making, and a right to opt out or withdraw. Further, the accompanying footnote 13 states that this "includes a right to individually or collectively decide how AI systems operate in a working environment", including provisions ensuring that "anyone using AI as part of his/her employment enjoys protection for maintaining their own decision-making capabilities and is not constrained by the use of an AI system". - With respect to "If one is a consumer or user of an AI system this entails a right to decide to be subject to direct or indirect AI decision making, a right to knowledge of direct or indirect interaction with AI systems, a right to opt out and a right of withdrawal": We agree with the principle that a user must be able to opt-out or withdraw from using an AI-based product or service. The AI HLEG may wish to clarify that if a consumer decides against using an AI-based product or service, non-AI-based alternatives may not be cost or operationally efficient for organisations. - With respect to footnote 13: This may suggest that employees are free to decide, individually or collectively, whether they can over-rule the AI system's recommendation. For example, if an employee objects to an AI algorithm making recommendations, despite the employer finding the AI algorithm able to make better credit or fraud prevention decisions, it does not seem appropriate for the employee to be permitted to ignore this. We support the AI HLEG's intention to safeguard against constraints on the decision-making capabilities of individuals, but would encourage the HLEG to consider the appropriate checks and balances for disregarding modelled outcomes.	Under the sub-section '1. Requirements of Trustworthy AI', we make the following comments: 1. Clause 2 - Data Governance. - With reference to "It is therefore advisable to always keep record of the data that is fed to the AI systems": We would like to highlight that it may not always be possible to retain all the data used to train an AI engine, or all the data that is consumed by the engine on a day-to-day basis. This is both due to practical reasons (e.g., volume of data that ever goes through the AI engine), and regulatory reasons (e.g., around data retention). We recommend that the AI HLEG clarifies that data must be retained according to applicable data laws and regulations. - With reference to "To trust the data gathering process, it must be ensured that such data will not be used against the individuals who provided the data": We would request clarity that this does not mean that AI decisions will always be in favour of individuals who provided the data (see Clause 5 below). 2. Clause 3. Design for all – We agree with the principle of this Clause. The AI HLEG may wish to reflect also that AI products and services are to be held to the same design standards as non-AI products and services. 3. Clause 5. Non-Discrimination – We agree with the principle outlined in this Clause. We would request that the AI HLEG also recognises the distinction between business decisions permitted by law, such as a decision to lend based on a borrower's factual financial records and credit history, and illegal discrimination. 4. Clause 8. Robustness: Reliability and Reproducibility – As the Guidelines recognise, the current state of AI does not necessarily lend itself to reproducibility. The AL HLEG may wish to consider the importance of materiality when examining the lack of reproducibility, as a greater degree of care and effort should apply to an AI product or service that has a greater material impact, for instance, having 'humans in the loop' in situations that can have a material impact on customers/staff/society. 5. Clause 9. Safety – We recommend that the Guidelines also reflect that AI implementation should not degrade an organisation's ability to meet its existing commitments and regulatory requirements around providing products and services that are accessible to all. With regard to the sub-section 'Technical methods', in general, the AL HLEG may wish to consider the importance of materiality when suggesting technical methods to achieve trustworthy AI, as a greater degree of care and effort should apply to an AI product or service that has a greater material impact. The Traceability and Auditability section mentions "Whenever an AI system has a significant impact on people's lives, laypersons should be able to understand the causality of the algorithmic decision-making process and how it is implemented by organisations that deploy the AI system."	Accountability – who is accountable - We would highlight that in practice, a number of different stakeholders or parties may be accountable when 'things go wrong'. Accountability - Diversity and Inclusion (D&I) – "Was a D&I policy considered in relation to recruitment and retention of staff working on AI to ensure diversity of background?". An organisation with a D&I policy should apply the same policy to staff working on AI; a separate D&I policy for the AI team should not be required. However, it is important to ensure that organisations carefully scrutinise AI outcomes in order to mitigate any possibility of unconscious bias. Accountability - Ethical AI review board – It is possible that not every organisation will require a separate 'Ethical AI review board'. Existing ethical and reputational risk forums inside organisations, duly empowered to consider AI, may be able to achieve the same desired outcome. This avoids creating additional overlap or fragmentation within organisations with additional committees or boards. Design for all – In line with our comment in response to question 3 above, AI implementation should not degrade an organisation's ability to meet its existing commitments and regulatory requirements around providing products and services that are accessible to all. Respect for Human Autonomy – The AI HLEG may wish to balance the ability of "users (to) have the facility to interrogate algorithmic decisions in order to fully understand their purpose, provenance, the data relied on, etc." against the need to prevent gaming of the system, and the usefulness of detailed technical explanations to laymen.	Glossary: We recognise that the Glossary provided is still incomplete, and will be further complemented in the final version. While there are no globally accepted definitions as yet in this space, we would encourage the AI HLEG to take into account international efforts as they emerge, to ease common adoption of the Guidelines and promote international convergence of standards, as well as provide any revisions or additional terms for further consultation before being finalised.

This principle may be balanced with two other considerations:
- The need to prevent gaming of the system: There may be a risk that individuals manipulate their data in order to achieve favourable outcomes. If AI is being used to assess fraud or financial crime risk, for example, a financial organisation will need to keep their algorithms private, as current regulations prevent similar transparency in human decision making (e.g., not tipping off a client on how to avoid getting caught in sanctions assessments).
- The practical difficulties of explaining the inner working of algorithms (even linear, rule-based ones): A broad explanation of the kind of data used, and the way in which AI is used to supplement human decision making, may be more useful and clear to those not familiar with the technology, instead of technical specificities.
For the avoidance of doubt, the above considerations that we outline do not apply to organisations' relationships with their regulators or obligations by law, which may demand transparency on both the data and the algorithms used to come up with decisions, as they do today (e.g., in validating regulatory capital calculation models).

With regards to the sub-section 'Non-Technical Methods', we support the non-technical methods proposed and agree with the AI HLEG that these should not be considered exhaustive or mandatory but instead used as a guide to help implementation. Considering the broad target audience for these Guidelines, which includes both public authorities and private organisations, the methods suggested are necessarily broad. Their importance will therefore vary depending on the stakeholders concerned. At a future stage, it may be useful for the AI HLEG to clarify which methods may be of more relevance to different types of stakeholders, to ensure complementarity and a more efficient implementation of trustworthy AI.

Under the sub-section 'Key Guidance for Realising Trustworthy AI', and with reference to "Ensure participation and inclusion of stakeholders in the design and development of the AI system": Stakeholders may widely refer to customers, employees, etc., and it may not always be feasible or practical to allow customers to participate in an AI system's design and development. Instead, the Guidelines could reflect "participation and diversity in the design and development of the AI system, and when setting up the teams developing, implementing and testing the product".

Bernd	Stahl	SHERPA, SIENNA, ORBIT, 4TU Centre for Ethics and Technology, Centre for Computing and Social Responsibility, Trilateral Research	About this response:This response represents the voices of the participants of two EU projects: SHERPA and SIENNA as well as the UK project ORBIT. SHERPA (Shaping the ethical dimensions of information technologies – a European perspective, www.project-sherpa.eu) explores ethical and human rights issues of smart information systems. SIENNA (Stakeholder-Informed Ethics for New technologies with high socio-ecoNomic and human rights impAct; http://www.sienna-project.eu/) is developing ethical protocols and codes for human genomics, human enhancement and AI & robotics, and the contributions to the text are from its AI/Robotics group. ORBIT (Observatory for Responsible Research and Innovation in ICT, www.orbit-rri.org) aims to develop a culture of responsible research and innovation across the ICT research community.The Centre for Computing and Social Responsibility at De Montfort University, Leicester, UK, Trilateral Research Ltd, and the 4TU. Centre for Ethics and Technology of the four technical universities in the Netherlands are among the partners who have contributed to this response.	The ethical principles listed in the document (section I.4) are well-established and have the advantage of forming the basis of processes of biomedical research. It is nevertheless surprising that the HLEG opted for adopting these principles and supplementing them with the principle of explicability, thereby firmly basing the approach to trustworthy AI on biomedicine. It is not obvious that this is the most appropriate approach, as it leaves out decades of research on ethics and computing or ethics and engineering, which may be equally or more relevant. One of the central conclusions of a recent FP7 project on ethical guidelines and assessments, SATORI, was that fields of computer science, engineering and social science involve largely different ethical issues from those in biomedicine and require their own ethical guidelines, distinct from those in biomedicine. We also believe that the granularity of the ethical principles is too low: only five ethical principles is not enough for a useful "ethical checklist" for AI, which would require a higher number of more specific principles.Having said that, we generally agree that the principles of autonomy, justice and explicability have applicability to AI, and constitute important ethical principles for this field. The principles of beneficence and nonmaleficence, however, have emerged historically within doctor-patient relationships and we find them too broad to be of use for AI. Both principles now function as a placeholder for a large number of more specific principles that relate to well-being, democracy, inclusiveness, fairness, mental autonomy, trust, sustainability, dignity, integrity, liberty, privacy, safety, security, and others. These are not, however, presented as sub-principles but are included in a larger narrative which is not systematized. We think it is better to "unpack" the principles of beneficence and nonmaleficence and replace them with four to eight more specific principles that are most important for AI. (Alternatively, a more structured list of sub-principles could be included under the headings of "beneficence" and "nonmaleficence"). A longer list of principles has the advantage that a separate "requirements" list will not be necessary; we think it is unnecessarily complicated to have guidelines that consist of five ethical principles which then translate into ten requirements.Good candidates for principles to replace these two principles are, in our view:Privacy: AI systems should protect (e.g., through privacy by design) and not harm privacy. Safety and security: AI systems should be safe for users and third parties, and should also provide security and resist being hacked or compromised.Well-being: AI systems should generally promote well-being and not cause harm to it.Responsibility/accountability: For AI systems that make decisions and perform actions that can cause harm or infringe on rights, there should be systems of accountability in place in which certain individuals or organisations are identified as responsible for the system's performance. Democracy: AI systems should generally promote and uphold democracy and not harm it; decisions that are normally made democratically should not be delegated to AI systems.Regarding the proposed critical concerns (section I.5), we agree with the importance of the first four. The fifth is now	The section on realising trustworthy AI starts with a set of requirements. While any of these requirements are reasonable and worth promoting, it is not clear how they relate to or are derived from the principles listed in the preceding section.Moreover, the items on the list are not consistent or commensurable. Data governance implies a set of well-established processes (e.g., the FAIR principles promoted by the EU) whereas robustness is a characteristic or a set of characteristics of the technology. It is not clear why these 10 items were chosen and not others. We also miss requirements relating to some of the principles of section I, including enhancement of well-being, respect for democratic procedures, and possibly prevention of misuse and dual use.As we suggested in our response to section I, we also think it is better to have one set of guidelines rather than two sets that are derived from each other. We discussed earlier how this might be accomplished. We believe that most of the ten requirements proposed in chapter II can double as ethical principles or guidelines, but two (data governance and governance of AI autonomy) qualify, in our opinion, as methods to achieve trustworthy AI and are more appropriately moved to section II.2. So, we propose to merge the ethical principles and requirements into one set of about ten ethical guidelines or requirements, and move the governance requirements to section II.2. Alternatively, the requirement of data governance could be renamed "data quality and integrity", which covers most of its current description, and then it could remain as an ethical guideline. In such a format, it should then include consideration for ethical issues which affect data quality and integrity and the relevant principles, including security, attribution and traceability, data minimisation, curation and retention, etc. and recognition that data governance requirements are not static throughout data lifecycles. Similarly, "governance of AI autonomy" could be reconceived as "human oversight of AI", in which case it has more resemblance to an ethical guideline than the current formulation.We suggest inclusion of a requirement regarding dual use and misuse. This follows from the principle of nonmaleficence, and correlated principles such as safety, security and well-being. This requirement is that AI systems should be designed and implemented in a way that anticipates and mitigates misuse and dual use. This can be a stand-alone requirement or it can be part of one of the broader principles or requirements.Where the requirements can come into conflict, there is no suggestion on how such conflict can be identified, addressed or resolved. The section on technical and non-technical methods is similarly not explained or derived from the principles or the requirements. What are the priorities and how are these to be implemented? Regarding the transparency requirement, there should be a discussion of the challenge that trade secrets and intellectual property rights pose to this principle and ways should be mentioned to overcome it. Otherwise, it only sounds like wishful thinking. This also applies to the corresponding explicability principle [p. 10] and traceability and auditability [p. 20.]Paragraph under "design for all" on page 15: In the first sentence, replace "citizens"	The section on assessing trustworthy AI repeats the requirements and provides some guiding questions on whether these are met. It is not clear what the status of the individual questions is and for what types of actors (Computer scientists? Corporations? Policy makers? Users?) these questions are intended. The introductory text stresses that the assessment is continuous and no step is conclusive. However, it is not clear whether all questions need to be addressed or what happens if some questions lead to answers that are contradictory while others seem to point to a requirement being fulfilled and not fulfilled at the same time? We advocate inclusion, under either Data Governance or Privacy, of a minimum data use principle for the use of personal data, as well as for mass surveillance. "No more personal data should be used for a task than is strictly necessary, and, further processing, storage and dissemination should be similarly minimized" (or the equivalent in question form).Under Privacy, shouldn't the first question be: does the system respect human privacy and have developers taken adequate measures to protect stakeholders' privacy?We suggest adding to Accountability: "Could the delegation of decision-making to the AI system allow individuals or organisations to unjustifiably claim diminished accountability for themselves for decisions made by means of the system?"We suggest adding to Governing AI autonomy the following item: "Is it ensured that the system is not made to make decisions that normally require human moral deliberation because they pertain to morally controversial decisions with significant impact, or democratic decision-making because they relate to common or public interests?" In addition, the item "What measures are taken to audit and remedy issues related to governing AI autonomy?" is very unclear and should be revised.Under Respect for Human Autonomy, second bullet: By "the latter", do you mean the service or product? Please check. Proposed revision: Has the developer or supplier provided users (stakeholders) useful and necessary information to enable the user to take a decision in full self-determination? Fourth bullet: Are we missing the word 'opportunity' here? I.e., do users have the chance and facility to interrogate algorithmic decisions? Under Robustness, third bullet: Why the sudden reference to 'my'? Under Safety, fourth bullet: "risk for" should be "risk to".We suggest adding to Design for all: "If usability testing is performed, does the test group sufficiently represent the diversity of the intended user base, including consideration for intersectional diversity along with gender, age, race, ability, education level and socioeconomic background?" In addition, the item "Is the system equitable in use?" should be made more simple or clarified further. How is a developer expected to answer this? Possible rephrase: "Is the system accessible for all users or stakeholders?" The item "For each measure of fairness applicable, how is it measured and assured?" is very confusing, consider revising. The fairness of user experiences should also be considered in addition to accessibility. Possible measures of fairness: Is it easy to find information about the system, its purposes, who to go to for more information. Is it accessible to all stakeholders? Accessible in terms of ease of use as well as cost? Is it discriminatory in	The document does a good job in summarising the debate around ethics and human rights in AI. It provides a conceptual basis and suggests requirements, actions, and evaluation. However, in addition to the lack of linkage between the different sections pointed to in the preceding parts of the response, the Draft Guidelines fail to address one key question, namely when ethical issues or human rights are sufficiently addressed to satisfy ethical criteria. This is a difficult question and aspects of it may be addressed in subsequent documents. There should nevertheless be a conceptual basis that allows for the identification of a level of ethical engagement that is sufficient, to facilitate practical work on AI to proceed. The current discussion of ethical and human rights implications of AI and related technologies shows a great amount of theoretical and conceptual insights, but suffers from a lack of empirical underpinnings. The HLEG might benefit from reaching out to research activities, both EU-funded and nationally funded that gather such empirical insights.An important practical consideration refers to one possible remedy mentioned in the Draft Guidelines, namely standardisation (p.21): There are a number of standardisation activities already under way (most prominently ISO and CEN standards, but also the IEEE standards group). We would like to point out that the EU project SHERPA already has a planned task and available resources for work on developing standardisation in AI. SHERPA would be happy to work with the HLEG and explore the potential for developing a relevant standard on ethics and AI. Some points regarding terminology and readability:We recommend using something other than 'North star' – this is slightly problematic terminology. "Pole star" would be better.The document is a bit hard to read at times and could be simplified. We recommend a review and revision of the use of language such as "It does not only" "avoiding to place" "receive greater attention to the prevention of harm", "Lastly, the principle of justice also commands .." "It should be born in mind" etcMore use of the active tense instead of the passive tense would improve readability and clarity. Replace 'sovereign intrusion' with 'state intrusion' or governmental intrusion'.Replace 'Healthcare Diagnose' with 'Healthcare Diagnosis' (p 28)

				formulated in vague terms, thus we cannot provide an evaluation or reflections upon it. Its reference to artificial consciousness does point to a critical concern that is more concrete: for the foreseeable future, AI systems cannot be morally responsible, have consciousness, have real emotions and pains. As a result, it would be wrong to depict them as having these properties, and in particular, to grant them the legal statuses that follow from them: personhood, rights, and citizenship.We also propose another critical concern, which is that AI systems should not be allowed to autonomously make decisions that go against the moral, legal and democratic order of society. Specifically, AI systems should not make any decisions that (1) are normally the subject of democratic decision-making procedures or stakeholder consultation (e.g., political decisions); (2) allow agents (humans or organizations) to delegate responsibility to AI systems and escape legal liability and accountability for their decisions; (3) normally require moral deliberation or conscience since they pertain to morally controversial decisions with significant impact; (4) go against prevailing legislation and regulations (unless defensible on the basis of an ulterior moral principle) or against widely accepted moral principles and norms.It is not clear how these critical concerns play a role in the formulation of the requirements in chapters II and III.On page 2, the second component of trustworthy AI should include security, which is critical to trust. So (2) it should be technically robust, reliable and secure. This is also exactly what is stated on page 17 under "robustness".	by "individuals", which is more inclusive, and add "gender" and "education level". Under "Diversity and inclusive design teams" on p. 22, the diversity dimensions of race and ethnicity should be included, as these are prominent in bias discussions. Other factors such as ability and socioeconomic status are also likely to be relevant, and consideration for the issues presented by intersectional bias should be incorporated.Additional technical/non-technical methods: Please consider adding the following on impact assessment:Impact assessmentImpact assessments are good tools for determining ethical, legal and societal impacts of an AI system, technology, product or service. Such impact assessments could be broad (encompassing all aspects) or specific (e.g., ethical impact assessment as outlined in the SATORI project/CEN Workshop Agreement Part 2: Ethical Impact Assessment Framework", CEN Workshop Agreement 17145, SATORI, May 2017; human rights impact assessment or a data protection impact assessment where legally mandated due to the existence of high-risk processing). An impact assessment that helps identify, assess and resolve adverse impacts of AI will boost transparency and build stakeholder trust. The Assessment List included in these Guidelines would form a part of the impact assessment exercise.	any way, against those of non-heteronormative sexualities, those with disabilities, women and non-binary individuals, ethnic minorities, religious groups, etc.We suggest adding to Non-discrimination, bullet 3: "...especially in the representation and reasoning about individuals and social groups". In addition, the question of the last bullet is too long and should be cut back or broken up. Or simply say: "Is it adequately clear to users or persons affected by the AI system to whom they can complain about any discrimination?"The Transparency requirement derives from the Explicability principle, which includes transparency of the basis of which AI systems arrive at decisions. However, in the bullets in this section, no explicit requirement is formulated for such transparency, only for transparency regarding the nature of the technology and potential risks. So, we believe that this requirement should be added here.	
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Anonymous	Anonymous	Anonymous		it is very important for me to describe clearly what informed consent is and the process of obtaining it. explicability is an important step in informed consent. for me, it is part of the principle of autonomy, so it is not necessary to isolate it. it is very important to take into account the technical skills of developers and business procedures (e. g. health) in the deployment of AIs in order to improve their quality and acceptability	Respect for (& Enhancement of) Human Autonomy: the obligation of users to tick the boxes of the system designer before having access is very important. The designer says he has the autorisation of the person to use his data but he has not given the choice: forced consent and not respecting the law in these cases.		It would be important to include all the actors likely to use the AI that will be developed. these people will be able to express their concerns, fears and expectations that will be directly taken into account in development. it is true that this is difficult, but this process will build confidence among all the actors

Patrick	Grant	BusinessEurope	The uptake of AI technology is highly relevant to business competitiveness and capability to innovate improved goods and services. Not least for addressing challenges in society like climate change, productivity and healthcare. The guidance on trustworthy AI could strengthen the uptake of AI, innovations and production, but then it must be relevant, meaningful and concrete. We also are also convinced that it is important that the guidelines strive to be maximising the benefits of AI as they are minimising its risks. AI will significantly transform (and is already transforming) the entire economy and society and will hugely impact the entire competitiveness of Europe and especially its companies, businesses, innovation and everyday lives. We do especially view the need for a balanced and future proof and coherent set up of the regulatory and ethical environment. Expand on the risk based and tailored approach to the AI initiative. Organisations should have the ability to tailor the approach to the impact on the individuals/society and risks related to AI application (within general principles of these guidelines). The guidelines could be used as a certification or signal to consumers that a company complies with the ethical guidelines. In that case it is of outmost importance that the guidelines are clear and concrete to ensure correct implementation by the companies. When guidelines like these are presented, companies tend to see them as mandatory and it is therefore important that the document is instructive and accessible. In conclusion, section II and III as such are far too extensive. Further regulation on AI and ethics may create unintended problems and limit the business ability/benefit.	Although use cases/examples discussed in this Section are pertinent, it would be advisable to refer to situations where consent is not achievable or can' be properly obtained because of the nature of the relationship, for example in an industrial workplace setting – the full possibilities of the GDPR should be possible (not contradicted). A right to decide to be subject (or not) to AI, a right to opt out and a right to withdraw significantly reduces the possibility to make use of AI systems. By definition, it relies on large volumes of retrospective data, making the execution of these rights impossible for any AI system, especially since typically AI systems will further use the input by users to improve the algorithms the AI system is built of. In addition, these requirements go beyond what it included in the GDPR, which regulates data protection. It is not the case that these additional requirements were omitted from the GDPR as it does have very specific requirements when it comes to automated decision making. It is not always possible to opt-out from a scoring mechanism without undermining the goals of the application of AI. It would be recommended to refer to a context specific application of such an opt-out mechanism instead. Security and Cybersecurity as a critical concern should be added to Section II 8 Robustness – p. 17, unauthorised access and manipulation of the systems with AI application raises new challenges from a security perspective as such interventions may not be as easily identified. A clarification might be useful – what does "a fair distribution of the value added being generated by technologies" actually mean? Is it fair access to the use and benefits of the technology that is intended (e.g. possible access to medical innovations based on AI), or is it a more "fair" distribution of funds/profits generated by the technology? Our presumption is that fair access is intended. More information on how to balance various freedoms with respective obligations and restrictions (i.e. freedom of the individual v national security obligations/cybersecurity restrictions). The paper might benefit from a bit more precision regarding the limits of "do no harm". International and domestic law is pretty clear that in certain situations (e.g. national security, to protect life and health, environment etc) there might be legitimate needs to overrule this principle. There certainly are cases where individual rights and liberties might be legitimately compromised in order to protect the rights, interests and liberties of others. Since these guidelines are intended to be an instrument that different stakeholders can endorse, it is important that as much clarity as possible is achieved to avoid disputes over what constitutes reasonable interpretations of the text. Identification without explicit consent is an existing, widely used practice not necessarily dependent on AI applications. The draft refers to GDPR article 6, which lists several legal bases for processing personal	What does it entail to endorse and sign up to the guidelines? More information needs to be taken on- board to explain the appliance to be compliant. The same goes for how updates will impact this. "…aims to reflect the main approaches that are recommended to implement trustworthy AI." – however some of the implementation methods read more as requirements, rather than truly offering guidance on how to practically implement them; for instance, in the case of "Explanation (XAI research)". Technical methods to achieve trustworthy AI are rather generic, and largely overlap with the principles set out at the start of chapter II.	What does it entail to endorse and sign up to the guidelines? More information needs to be taken on- board to explain the appliance to be compliant. The same goes for how updates will impact this. "…aims to reflect the main approaches that are recommended to implement trustworthy AI." – however some of the implementation methods read more as requirements, rather than truly offering guidance on how to practically implement them; for instance, in the case of "Explanation (XAI research)". Technical methods to achieve trustworthy AI are rather generic, and largely overlap with the principles set out at the start of chapter II.	The sign-up creates one of the biggest concerns. What does it entail to endorse and sign up to the Guidelines? The guidance should be more focused and shorter for ease of reading and to gauge the interest of those outside Brussels. The scope of application of requirements (technical and non-technical) should be linked to the risks and impact on the individuals. AI is being used in many industrial applications (eg. automatically steer drilling operations and use of AI to predict maintenance priorities (predictive maintenance). Another use case is use of AI to analyse CCTV footage to detect health & safety risks (eg. smoking in forecourts of petrol stations). These industrial use cases are somewhat underrepresented in the guidelines and we recommend further focus on this type of application. Trustworthy AI includes considering cybersecurity risk and, in the context, especially of AI applications making use of IOT devices for example this needs to at the forefront of policy. The guidelines could be more detailed and explicit in that respect. Trustworthy AI should consider known threats to interference with the AI system and ensure that they are mitigated to the fullest extent possible. All in all, it would be essential to assess the impact of how the approach with ethical purpose will affect innovation in Europe, especially the competitiveness of companies focusing on AI. It should be remembered that the industry does not want or need extra barriers to business, which is one of the cornerstones of European welfare.

					information. Picking out consent as the primary justification is unfounded and goes against the technology neutral approach of data protection. Identification and (other) processing of personal data should be allowed on any lawful grounds recognised by GDPR or other relevant legislation. The wording of the principle of autonomy is too far-reaching. Already today consumers, workers and other users are subject to automated decision making, whether based on AI or simpler applications. E.g. the right to opt out could conflict with existing employee obligations and lead to dismissals. The principle of autonomy should be limited to freedom from coercion not everything else. "AI systems should be designed and developed to improve individual and collective wellbeing (…) by generating prosperity, value creation and wealth maximization". This sounds very generic. What does this mean specifically? Any company making more money thanks to AI is in accordance with this principle?		
Gabriele	Trovato	Waseda University				Dear all,together with my students in the AI class, we have discussed the guidelines. I report here a few points that are worth mentioning. I hope they might be helpful.(Credit to Kaho Ko for raising the original discussion point)The guidelines stress the importance of agreement of the user in the use of one's own information. However, it does not put emphasis on the punishment or action to be taken if the AI or technology system breaches the agreement. What would be the counter action if the AI does not comply to enstablished ethical values?We know that users typically give consent without paying much attention to the content. Big companies can make profit and are unlikely to follow rules.As these guidelines are not enforced by law, I would rather propose the realisation of a AI-related CE Marking for European Conformity, which guarantees the quality of an AI in an easily understandable manner.(Credit to Hyunah Kang for raising the original discussion point)Regarding ensuring that users are always aware that they are interacting with and AI rather than a human.The goal of developing human-like AI and not allowing it to be covert contradict with each other. Rather than hindering the development of AI applications and of the world wide research on androids, we should regulate in which cases it's okay to be covert. (Credit to Raoul Man for raising the original discussion point)On the Principle of Autonomy: "Preserve Human Agency":While users/consumers of AI systems are mentioned, there's little to no mention about those indirectly affected by someone using an AI system, for ex. AI driven cars and pedestrians, other drivers in traffic. How do you ensure that those that, at that point, become part of an AI's use are first of all aware that an AI is being used and second of all allowed to opt out of the use of that AI?	

Carl	Wiper	Information Commissioner's Office	As the regulator for information rights and data protection in the UK, and current Chair of the International Conference of Data Protection and Privacy Commissioners (ICDPPC), the ICO welcomes the work of the High Level Expert Group and the opportunity to respond to the working document on draft ethics guidelines for Trustworthy AI. In doing so, the Information Commissioner recognises the importance of a shared ethical framework underpinning the international landscape of AI governance, building on the Declaration on Ethics and Data Protection in AI agreed at last year's ICDPPC conference. We support the identification of Trustworthy AI as the 'north star' of the High-Level Expert Group, and particularly the requirement that AI be 'demonstrably worthy of trust.' The ICO has found evidence of low levels of trust by the public in how organisations use personal data and this represents a potential barrier to the development of AI. It is only when data controllers and processors are in a position to demonstrate that they are worthy of the trust that may be placed in them that the benefits of AI can be fully and ethically realised. This chapter also references a future mechanism to enable stakeholders to sign up to the guidelines, and the ICO would be interested to learn more about the role such a mechanism is expected to play. We welcome a greater degree of co-ordination and co-operation in this space, recognising the connection between digital ethics, governance and regulation in the realisation of trustworthy AI.	The ICO welcomes the rights-based approach to AI Ethics in the draft guidelines (which complements the protection of human rights and freedoms operationalised in the GDPR) and the derivation and development of ethical principles from these rights, These help to reinforce key data protection principles, as recommended in the ICO's 2017 paper 'Big Data, Artificial Intelligence, Machine Learning and Data Protection'. Section 3.4 on equality talks of "inclusion of minorities, traditionally excluded, especially workers and consumers" (p7). It seems a little unusual to categorise workers and consumers in this way. We would recognise that it is possible to have an imbalance of power between workers and consumers on the one hand and data controllers on the other, but it does not seem correct to refer to these groups as minorities. The introductory paragraphs to section 4 (p8) advise "the presence of an internal and external (ethical) expert". Some organisations deploying AI will have limited resources and the guidelines should present an approach which is scalable to their needs. This statement could therefore perhaps be qualified with a phrase such as 'wherever practicable'. The ethical principles articulated in section 4 reflect those used in the field of bioethics. We make no comment on how successfully they have been implemented in that field, but we do note the addition of a fifth principle: explicability. It could be argued that explicability would not be universally recognised as a normative principle, but nevertheless we think it is appropriate to add it, provided it is interpreted broadly to include concepts such as explainability, intelligibility, transparency and accountability. We see it as a principle which can enable the application of the other principles and provide an assurance that they are being followed. There are also important linkages with the GDPR principles of transparency and accountability, and the GDPR requirements for meaningful explanation of automated decision-making. In this context, the ICO is currently working with the UK's Alan Turing Institute on producing guidance to assist organisations in explaining decisions made by AI systems.	The explanation of accountability in the list of ten requirements for realising trustworthy AI seems to focus on mechanisms for compensating for error or wrong-doing, rather than pro-actively ensuring compliance. This doesn't cohere with the meaning of accountability in the GDPR, where it is understood in a wider sense as being responsible for, and able to demonstrate compliance with, the data protection principles. It may not be helpful to use the same term in a more limited way in the ethical guidelines, and we would prefer it to be used in a wider sense here. Regarding non-discrimination (section 5), it may be worth distinguishing between two kinds of unintentional bias in data. The first is the bias that arises when the data is not drawn from a statistically representative sample of the population of interest (e.g. containing proportionately fewer women), which results in less accurate models. The second kind of bias concerns data which accurately represents the population (e.g. containing a proportionate number of each gender), but where this in turn reflects the results of direct or structural discrimination (e.g. workplace assessments which reflect the gender biases of managers or unfair maternity arrangements). Reference to the GDPR in the list of requirements for trustworthy AI is limited to the requirement for Respect for Privacy, but the scope of the GDPR extends to a number of the other categories, and it may be helpful to acknowledge this. In addition to accountability, there is a clear requirement for transparency (1st data protection principle), non-discrimination (e.g. recital 75), robustness (accuracy, 4th data protection principle; controller obligations, e.g. Articles 25, 35). These are legal requirements for data processing under the GDPR as well as ethical requirements. It would be helpful if reference to compliance with the GDPR were not limited only to the requirement for Respect for Privacy. Following on from this, some of the technical and non-technical methods for achieving trustworthy AI align with the legal requirements under the GDPR for data controllers and processors to ensure that data protection principles and data subject rights are complied with by design and by default. In particular, a Data Protection Impact Assessment (DPIA) is likely to be a requirement for AI projects processing personal data. The ICO would expect UK data controllers to demonstrate their compliance as part of a DPIA using at least some of these methods, as part of an ongoing process and in line with their legal obligations. Section 2.1 Testing & Validating argues that 'bounty hunting' may be considered, whenever feasible, as a technical method of achieving Trustworthy AI. In the context of the GDPR, this may not be advisable. Depending on how the 'bug bounty' program is organised, vulnerabilities exposed by even 'white hat' hackers may still constitute data breaches that need to be reported to the respective data protection authority (DPA).	The draft Assessment List for trustworthy AI, although not proposed as mandatory for AI developers and companies, would sit alongside the likely legal requirement for a DPIA under the GDPR. We would expect a number of the questions on the list to be addressed by data controllers as part of a DPIA for an AI project, so there is a broader question here about how these assessments sit alongside one another – or might be brought under a single form of assessment – where the appetite among some organisations to carry out two discrete assessments may be low. Outside the GDPR jurisdiction, the Ethical Data Impact Assessment model developed for the Hong Kong Privacy Commissioner is an interesting example of bringing together ethical and data protection assessments. Under the assessment questions for Respect for (& Enhancement of) Human Autonomy (p26), there is no reference to consideration of a right to opt out or withdraw from AI systems and decision making, although these rights are mentioned under the principle of autonomy in Chapter I. Perhaps this ought to be part of the Assessment List when considering autonomy, and this may lead to a consideration of how such rights can be actualised, given how pervasive AI systems and decision-making may become. The ICO is interested to understand the process by which such assessments would be carried out, with the inclusion of "specific metrics" (p.24), and looks forward to learning more in the next iteration of this document.	The ICO welcomes the HLEG's framework for trustworthy AI, as comprising ethical purpose and the need to be technically robust. Europe is already a global leader in the regulation of information rights, reinforced by the introduction of the GDPR, and additional compliance with agreed ethical guidelines will help further position Europe and the UK to reap the benefits of AI. We are keen to see the realisation of these benefits, both for individuals and for society as a whole, encouraging innovation that is compliant with data protection law and with people's rights and freedoms. Having said that, the statement in the Executive Summary, that "on the whole, the benefits of AI outweigh its risks" seems rather too generalised to be meaningful. It may be better to make a statement to the effect that AI can bring enormous benefits to society and to individuals, but its development requires a respect for fundamental rights. We believe that a drive towards the ethical development of AI will serve to support the work of DPAs in protecting personal data rights. Assessing fairness in the context of AI increasingly raises issues to do with the societal impacts of the processing which are difficult to resolve within the scope of data protection legislation, and the development of ethical standards can help there. Furthermore, the adoption by organisations of ethical approaches to data use will serve to assist their compliance with legal requirements. The ICO is willing, within the limits of its remit, to contribute to the development of this framework, working together with partners in data ethics to "develop a unique brand of AI" (p ii). In the UK the creation of the national Centre for Data Ethics and Innovation is an important step in the same direction and we are planning to work closely with them. We recognise that these draft guidelines aim to foster reflection and are a starting point for discussion on 'trustworthy AI made in Europe'. While our remit is to be the regulator for the data protection laws that protect UK citizens, the development of a wider international consensus on 'the common good' in relation to AI technologies would be a welcome result of such discussions.

Anonymous	Anonymous	Anonymous		I 3.5 This section only addresses citizen-government relation. I propose to extend the coverage to private sector ie corporates as well. Citizens should have rights when dealing with private corporates that apply AI, e.g. the right to know if their data is handled by automated systems and the right to opt out. 5 Let me propose three additional concerns that I think are critical. i. Impact on individuals or group of individuals thinking, world view, conscience. AI systems may be used to impact unintentionally, like creating echo chambers/filter bubbles around individuals, or intentional, like persuasion or manipulation. In both cases AI systems can be extremely efficient and unnoticed by the individual, thus we need proper ethics principles and also regulation that informs the citizens and prevents damage. Well known example is Facebook's activity in creating echo chambers, manipulating political opinions and elections, persuading spending decisions. ii. Purpose alignment and containment of AI systems. This is covered in the 'Safety' section of the upcoming chapters, I'm proposing to add these issues as Critical concerns that require topmost attention in regulation. iii. Data security issue on national and also on European level. Foreign/international corporations have huge amount of data on citizens of individual countries and also Europe. Leading social networking corporations can easily have better information about a country's citizen then the government itself. Without proper regulation in place, they may use this extremely valuable information for rouge purposes.	II 2 I propose to add the data security issue on national/European level, ref. my comment on Critical concerns iii. Pls ensure that corresponding questions/challenges are added to chapter III. 6 I propose to extend this section with issues as follows - Citizen should know if communicating with/treated by AI - Citizen should be able to opt out any time form being treated by AI - Unintentional or intentional persuasion/manipulation of citizens should be avoided Pls ensure that corresponding questions/challenges are added to chapter III. 9 I propose to add here the issue of containment, proper container must be in place to restrict unintended consequences. Pls ensure that corresponding questions/challenges are added to chapter III.

andrea	simoncini	University of Florence	(P. 2) "Purpose and Target Audience of the Guidelines"After "-companies, organizations, ...other entities" - I would specify "academic institutions" or "educational agencies" in order to emphasize (as after will be clarified in the Guidelines) how decisive is the role of education and training of AI specialists for internalizing the aims and values of these Guidelines	(P. 5) Proposal 1 The EU's Rights' Based Approach to AI Ethics- in the footnote (1), I would cite also art. 4 TEU-after "and promote the common good" I would add "respecting Member States' constitutional identities"Here is the text of Art.4 TEUArt. 4.1 The Union shall respect the equality of Member States before the Treaties as well as their national identities, inherent in their fundamental structures, political and constitutional, inclusive of regional and local self-governmentMotivation:The reason for this proposal is to acknowledge that the European Union is a "plural constitutional entity", where different levels and standards of constitutional protection of rights are guaranteed and sometimes they may be higher than European ones. AI, in those cases, has to comply with the higher levels of protection.Proposal 2From Fundamental rights to Principles and ValuesAfter "In turn," I would add "effective and fully aware" before "informed consent"Motivation:We all know (and the Guidelines also tackle the issue) how often the consent is too weak protection against the asymmetry between the technological service needed and the individual freedom.We have to insist on qualifying the "informed consent" as "effective and fully aware"P. (7)Proposal 33.5 Citizens rightsI'm aware in this paragraph you're referring to EU Charter but (here is a case in which the EU HR protection standard may be lower than national constitutions) why to restrict those rights only to "citizens in their interaction with the public sector"? So excluding all non-citizens (as foreigners or migrants) and citizens in their interaction with private companies.Why not-citizens should not have the right to a good administration, to access to public documents, and the right to petition the administrationWhy citizens shouldn't enjoy the right to be informed of any automated treatment of their data by PRIVATE bodies and systematically be offered to express opt-out (GDPR does)Why citizens should not be subject to systematic scoring by the government while private companies can?(P. 10)The Principle of Explicability: "Operate transparently"After "Individuals and groups may request evidence of the baseline parameters and instructions given as inputs for AI decision making (the discovery or prediction sought by an AI system or the factors involved in the discovery or prediction made) by the organisations and developers of an AI system, the technology implementers, or another party in the supply chain"I would add that this right will prevail against any kind of copyright or intellectual property right. Otherwise, the above paragraph is totally ineffective when algorithms are considered under the protection of intellectual propertyWhen fundamental rights are concerned, the right to explicability has to prevail on any conflicting right.			It could be useful within Part II, Non-discrimination principle (P.16)trying to be more detailed in specifying which measures or rules have to be followed to avoid biases in the input datasets in order to avoid algorithmic discriminations.-
Rebecca	Jungwirth	F.Hoffmann-La Roche	Roche welcomes the drive to frame concerns and expectations around AI. Such a document is needed, and we will be happy to engage in the debate going forward.		The ten requirements of trustworthy AI are a very good basis for discussion. Teasing out the various aspects and interactions of data vs. system vs. human element is challenging, and as a result there is a certain amount of overlap between sections, and it is at times challenging to differentiate aspects from one another. To make it easier to follow, further attempts to clearly	The listed questions highlight useful areas of consideration. What is, however, often missing is a reference to what is considered the current gold standard that an AI would seek to emulate. In other words, it is not always clear what to compare AI to. In addition, some of the questions listed in Section III imply that AI always requires	Overall, the document could benefit from a clearer structure to prevent repetition and overlaps between multiple concepts and ideas throughout. This would allow for a clearer presentation of the main points.

				separate the ideas would help. The brief segment related to hacking offers worthwhile considerations, and could be expanded.	consent, e.g. “How can users seek information about valid consent and how can such consent be revoked?” Depending on the type of technology and its application, consent may not be required; the GDPR (and privacy laws in general) concern lawfulness of data processing, not consent. This could be easily rectified by changing the question to, “Provided consent builds the legal basis for the processing of identifiable information in an AI application, how can users seek information about valid consent and how can such consent be revoked?”	
Simon	Foley		We welcome the ongoing theme in the report that a human-centric approach to AI is needed which places the autonomy of the individual at the centre. Informed consent forms an important part of helping to achieve autonomy of the individual and we would like to see greater prominence given to the ethical challenges surrounding informed consent in AI-based systems in the guidelines. For instance, Sections 5 considers identification without consent (5.1), however ethical practices surrounding the use of anonymised data and consent is not considered. In this case, one viewpoint is that the collection and use of anonymised data in an AI system, without the informed consent of the data giver for its particular use, is acceptable if it increases human well-being. However, another viewpoint is that this utilitarian ethical view may be at variance with the autonomy of the individual: the individual may have principled reasons and/or beliefs for not consenting to the use of their (anonymised) data in an AI system. Informed consent, as described by the GDPR, makes significant strides on these challenges at a regulatory level. We think that in these ethical guidelines for trustworthy AI there is an opportunity to strengthen the ethical practices with respect to achieving autonomy for the individual. We believe that much of this ethical purpose is best framed in terms of virtue ethics, that puts the autonomy and the well-being of the individual to the fore. Related to this, on Page 10 the report notes that "Humans might benefit from procedures enabling the benchmarking of AI performance with (ethical) expectations". We agree: benchmarking can provide a means for organisations to understand and communicate their own ethical practices in the use of AI and also provides them with a trajectory for improvement. We envisage levels corresponding to different ethical practices, such as utilitarian ethics -> legalistic ethics -> ethics of virtue. Such benchmarking enables parties to subscribe to guidelines through ethical profiles that best match their own practices and ideals. In our own work we have suggested these levels in an ethical maturity model for informed consent [1]. We would also suggest that [on Page i] the statement: "human-centric approach to AI is needed, forcing us to keep in mind that the development and use of AI should not be seen as a means in itself, but as having the goal to increase human well-being" conclude with "[...] while being mindful of the desire for the autonomy of the individual".	Section 2 stresses that achieving trustworthy AI is a continual process. We note that this is consistent with the practice of ethics of virtue and is welcome. It could be argued that the technical method "Ethics & Rule of law by design (X-by-design)" promotes design thinking as "design for use before use". A challenge with this kind of design thinking is that it may encourage a legalistic ethical view (Rule of law by design) whereby the rules governing the AI system are considered elicited before the system is in use. However, there is the potential for a "Symmetry of Ignorance" between an AI system designer and the AI system user: before use, the AI system designer may not properly understand the end-user's desire for autonomy in the AI system nor appreciate how their system may impact it. On the other hand, the end-user may not properly understand the designer's objective for the AI system (nor indeed their own desires for autonomy). This Symmetry of Ignorance is evident in the development of security systems [3]. Thus, a Trustworthy AI architecture should also recognize that it is only through using the system that certain ethical issues may come to light: premises about system requirements and user needs change as designer and end-user understanding increases. This is suggested by the process in Figure 3 in the guidelines, although it should be emphasised in the guidelines that the design thinking for Trustworthy AI should be "design for design after design" rather than "design for use before use". This also reflects how we develop and use contemporary systems: user needs evolve, requirements change and new technologies are incorporated, all as a part of the normal 'use' of the system. Thus, it is not possible to design such open systems that anticipate all uses and ethical challenges in advance and therefore one must “design to support design after design”. We welcome the guidelines on respecting and enhancing human autonomy (Section 6) and on Privacy (Section 7). In our own research on data ethics and informed consent, we have found it worthwhile to draw from the ethical practices that have evolved in Qualitative Longitudinal Research techniques that are used in Applied Psychology and Social Sciences. We believe that these ethical practices can also help provide guidance on realising trustworthy AI. Resources from the tradition of Qualitative Research, particularly longitudinal research, such as reflexivity and relational ethics are tools that can be applied in order to enhance ethical practice. In longitudinal qualitative research, where participant and researcher have repeated contact for data gathering, the issue of consent can be revisited during	Gathering and acting on information from people is not a recent phenomena, nor is it unique to AI settings. For decades, psychologists and social scientists have studied human behaviour. This has involved information being gathered from, and about, ordinary people. Similar to AI systems, psychological studies gather data over time, retain the data, and subject it ongoing analysis. This is especially the case with Qualitative Longitudinal Research (QLR). The similarity between QLR and AI systems is in obtaining, retaining, analysing and acting on information about people. Another similarity is the nature of the information disclosed, as personal, and often sensitive information, is revealed by people about themselves. During collection and analysis, such information can be linked to previous occasions of data gathering, and made sense of in the context of our cultural and social world. Inferences are drawn about people based on scrutiny of what has been disclosed and retained. Shared ethical issues arise in both contexts, in light of the uncertainty of outcome following analysis, and the consequences for individuals.Over time, in the practice of QLR, a body of theory and practice on informed consent has evolved, such that ethical conduct can be fostered. An ethos of evolving ethical practice underpins the mature approach to informed consent that has emerged, and this enables researchers to respond to new ethical dilemmas that necessarily arise in practice. Our position is that how and why informed consent has developed in QLR can be used as a resource to inform the development of a similar process for AI systems. How informed consent has evolved in QLR can be characterised in terms of a a three level maturity model [1], based on ethical approaches that are (1) Utilitarian, (2) Principled and (3) an Ethics of Virtue. The model illustrates how Informed Consent in QLR has progressed over time toward the ideal approach of an Ethics of Virtue. Given the shared characteristics in the contexts outlined above, we argue that the model provides a means to assess informed consent in AI systems, enhancing development of best practice, and creating an ethos of fostering ethical practice. In practice this could mean, for example, that individuals are regarded as stakeholders in any data held about them, and as such, are consulted about its use. In addition, as the maturity model affords the opportunity to analyse practice we can, therefore, shed light on the theoretical underpinning of any given instance of Informed Consent. We can learn whether an approach is appropriate in light of the relationship between the parties when consent is being sought. For instance,

						such interactions, as well as other occasions, such as when data analysis is being finalised, or questions around participation arise. At such times, decisions pertinent to the analysis being conducted, or any other aspect of the research, such as its purpose and the potential use of the data, can be the subject of reflexivity on the part of the researcher, and this includes discussion with, and input from, participants. The right of a participant to withdraw their data, and to revoke their consent to participation in research, can be reiterated and discussed at these times. The objective is ensuring that consent is truly informed, that is, that the participants are aware of their autonomy in the process, and that this is made apparent to them by the researcher. Another example of this approach is where there is an interval between the request for consent being made by a researcher, and any possibility of agreement to the request by a potential participant, facilitating reflection on what is being requested, and facilitating an easy for a participant to choose not to consent. Approaching the issue of informed consent from this more participatory ethos makes for a more equitable research relationship [3]. The benefit of this approach is in fostering mutual trust between participant and researcher, as power is ceded by the researcher by ensuring that participants are aware of their own power as research participants, and that the process of informed consent is conceived practically and ideally as a device to highlight and achieve this awareness. Furthermore, the approach also creates awareness that unknown moral dilemmas will arise concerning the use of data, and that resolving such dilemmas is a process that is best achieved from a participatory ethos, encompassing self scrutiny. Self scrutiny can be achieved by a researcher choosing to act with empathy when decisions are being made. These resources from the tradition of qualitative research are envisaged as the basis for developing reflexivity in AI. In summary, reflexivity means that the researcher scrutinises their own actions, and that a positive encouragement to behave ethically is adopted, rather than focussing on prohibitions. We therefore recommend the use of reflexivity as a further non-technical method (Section 2) that can help in realising trustworthy AI.		if a power or knowledge disparity exists between the parties, then such an unequal relationship can mean that the proposed agreement may be weighted in favour of one party at the expense of the other. By adopting a theoretical approach to the analysis of informed consent, the possibility of fostering an ideal of ethical practice becomes part of the discourse. Including theory in the discourse around Informed Consent demonstrates our societal aspiration to ensure that fairness permeates the concept of Informed Consent. In the absence of such an aspiration, fostering ethical practice will remain a challenge. Simon Foley & Vivien Rooney----------References[1] VM Rooney, SN Foley. "An online consent maturity model: moving from acceptable use towards ethical practice" In New Security Paradigms Workshop (NSPW 2018). ACM press. https://arxiv.org/abs/1710.10022[2] VM Rooney, "Consent in longitudinal intimacy research: adjusting formal procedure as a means of enhancing reflexivity in ethically important decisions", Qualitative Research Journal 15(1), pp71-84 2015. [3] O. Pieczul, SN Foley, M-E Zurko "Developer-centered security and the symmetry of ignorance". New security paradigms workshop (NSPW 2017), pp 46-56, ACM Press.
Johannes	Holtz	DATEV eG	We welcome the first draft of the “Ethics Guidelines for Trustworthy AI” by the EU-Commission’s High-Level Expert Group. For us as an IT service provider, data security and data protection have the highest priority and are of fundamental importance. DATEV stands for exceptionally high standards in this area. We provide the software for 40.000 tax advisors, for the financial accounting of 2.5 million enterprises (mostly SMEs) and for around 13 million wage and salary statements per month. This data falls under different, often very high levels of confidentiality. Therefore, we welcome the approach to put the trustworthiness of AI in the center of the European approach. AI made in Europe must be trustworthy, not only in order to achieve the greatest possible benefit for society, but also because it is our greatest advantage in order to prevail in international competition. For Trustworthy AI			We welcome the assessment list for trustworthy AI as well as the plan to adapt the assessment list to four uses cases: (1) Healthcare Diagnose and Treatment, (2) Autonomous Driving/Moving, (3) Insurance Premiums and (4) Profiling and law enforcement. We would propose to add a fifth use case: taxation and finance. The potential for AI in this area is large. Hence, a specific assessment list for the trustworthiness of an AI is necessary.	For a successful adoption of the guidelines by stakeholders it is of great importance to make the guidelines available in different languages, preferably already during the design phase.

			made in Europe, European AI Ethics Guidelines are a first fundamental step. We have noted with great interest the mechanism to be put in place enabling all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis.					
Florian	Baltruschat	German Insurance Association (GDV)		Regarding AI use by the insurance industry, we believe that responsible and trustworthy AI is already ensured: Insurers are subject to a comprehensive regulatory and supervisory framework. The insurance supervisory authorities (EIOPA, BaFin) closely monitor and supervise insurers' AI usages, and the insurance industry is very experienced in using data and new technologies in a responsible and secure way. Regulations of the analogue world (e.g. information requirements) automatically apply to the digital world as well. In addition consumers' right regarding data protection and automated individual decision-making were strengthened by the EU General Data Protection Regulation (GDPR). For instance, pursuant to Article 13(2)(f) of the GDPR, consumers shall be informed about the use of automated individual decision-making, including meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject. We agree with the HLEG's assessment of the potentially huge benefits of AI for society and human well-being. In order to exploit these benefits, it is crucial that the framework regarding AI use (e.g. regulatory provisions, ethical guidelines) not only limits risks and safeguards fundamental rights, principles and values, but at the same time is innovation-friendly and does not hamper effective competition and companies' endeavors to find better solutions for their customers. Overly restrictive requirements should be avoided and the guidelines should be interpreted carefully. For example, that a company's changes in product design (e.g. replacing some features), customer service (e.g. available communication channels) or price system often benefits some customers while other customers are unfavorable impacted is a commonplace occurrence in a market economy and a driver of the competitive process when customers search for better offers. This distributive effect should not be interpreted as unfairly harming some customers. In the current discussion on the regulatory framework regarding FinTech, technology neutrality and the principle of proportionality have been identified as fundamental regulatory principles. The ethical guidelines for trustworthy AI should be consistent with these regulatory principles. Requirements should be proportionate to the risks involved, irrespective of the technology used. We very much support the HLEG's context-dependent approach and the adaptation of requirements to concrete use cases. In particular, the very different risks and circumstances of the manifold uses of (weak) AI should consistently be taken into account. It is important that the guidelines provide sufficient scope for interpretation to ensure appropriate solutions for the different use cases.	Regarding the use case "Insurance Premiums" it is crucial that the proposed requirements are adequately interpreted and adapted in order to take the characteristics of insurance products and the prerequisites of effective insurance markets sufficiently into account. In particular, the interpretation of fairness should be based on the principle of risk-based pricing that is fundamental for effective insurance markets, reliable insurance cover for customers and the financial stability of insurers. We therefore encourage the authors of the guidelines to clarify that equal treatment of all human beings does not imply equal prices for all human beings. This is in line with the European Court of Justice, which has consistently held the principle of equal treatment requires that comparable situations must not be treated differently, and different situations must not be treated in the same way, unless such treatment is objectively justified. In private insurance, every customer pays a premium based on the risk that this person brings into the pool of insured. Equality and fairness in insurance means, that this principle is applied to all people in the same way. This principle results in different prices for people with different risks and consequently equal prices for people with equal risks. It has to be noted that in different settings, e. g. when insurance cover is compulsory or when assessing fairness of products of other industries, other approaches to equality could also be valid. The different approaches to equality and fairness should be addressed by a context-dependent interpretation of these terms. Applying different fairness concepts simultaneously to a certain application in search of solutions that are unambiguously fair for all is not appropriate. Even with new efforts to measure fairness of AI-applications with mathematical formulas, improving the applications‘ fairness with regard to one concept often leads to poorer results with regard to other fairness-concepts.	We appreciate the opportunity to participate in the consultation of the current draft. We also kindly request the expert group to hold another consultation on the final draft: Since two of the upcoming assessment lists planned use cases (insurance premiums, autonomous driving / moving) will revolve around insurance and therefore are of high importance for the insurance industry. In addition, a third use cases on (Healthcare Diagnose and Treatment) has a substantial indirect impact on the insurance industry. However, a substantiated assessment of the operationalisation of the assessment list will only be possible on the basis of draft lists for the two use cases.

Ian	McArdle	Communications Workers' Union - Ireland	Europa welcomes the possibility to contribute to the stakeholders' consultation and underlines the importance of a broad public debate and information on AI. This debate must result F1+G286+F1in clear ethical and social guidelines and standards with the aim of improving the living and working conditions of European citizens.- We acknowledge the innovative potential of AI and new technologies that can be beneficial for our society. However, these new technologies also create challenges and we are concerned about the possible risks and consequences relating to working conditions, skills and training, ethics, equality, health and safety (among others). Therefore, UNI Europa would like to underline the importance of addressing AI technologies and robotization as topics for collective bargaining at all levels (company, national and European). AI and robotics have a huge impact on the future labour market, as jobs will sometimes disappear or be transformed and other jobs will be created. We need to accompany this process and address the question of skills and training for the future workforce: need to ensure that training on necessary digital skills is provided by education institutions and companies, and that it is not the sole responsibility of the worker to keep up with the rapid technological developments. Employability needs to be promoted through upskilling and reskilling schemes for workers. Investment in formal, informal and life-long learning is key; we must enable people to work with AI or invest in competences that AI will not cover. It is important to develop action plans at EU and national level together with education providers and social partners in order to modernize education and vocational training. We therefore welcome the call from the ILO Global Commission on the Future of Work for "a formal recognition of a universal entitlement to lifelong learning and the establishment of an effective lifelong learning system. (https://www.ilo.org/global/topics/future-of-work/publications/WCMS_662410/lang--en/index.htm)- The social partners play a key role in this and the EU should cooperate with them and national governments in order to identify which job sectors will be affected by AI. We need to understand the timeline and extent of changes in the labour market. The involvement of social partners is a must to find appropriate and future-proof solutions to concerns relating to employment, training, the nature of work, (in)equality or social systems and collective bargaining, especially at sectoral level.- As AI and automation have the potential to transform not only simple tasks but very complex processes, we need to have a large public discussion about the areas in which the use of AI is reasonable and beneficial for society. Part of the debate should be the question of how the profits generated by AI should be re-invested for the common good by creating employment in domains such as care, health services, education or mobility. Employees should participate in the distribution of profits, e.g. through wage increases or reduction of working time. Moreover, AI wins should be used to strengthen social security systems. This could be a measure to address the problem of future job losses and the precarisation of employment relations in a-typical work (e.g. platform work) due to AI and automation.- It	- UNI Europa supports the human-centric or human-in-command approach suggested in the guidelines. We agree that it is necessary that humans always need to remain in control of technology and machines. Likewise, we agree that the use of AI needs to respect European values and fundamental rights.- We recommend an expansion of 3.2 "In an AI context, freedom of the individual requires protection from direct or indirect coercion, surveillance, deception or manipulation." Much profiling that lies at the heart of AI systems relies on a degree of manipulation of data. This is not least relevant in relation to the world of work and especially in the use of AI in Human Resources. - We welcome that the HLEG understands the need to ensure that those involved in the development and marketing of AI (researchers, engineers, designers etc.) act in accordance with ethic and social responsibility criteria. This should be addressed by changing educational priorities for technical subjects and by providing lifelong learning opportunities (e.g. by incorporating ethics and the humanities into training courses in engineering). - Organisations and companies should develop tools to facilitate ethical discussions and decision-making throughout the whole design process. This should be completed by internal training programs on ethics for all employees. Such training should help employees understand the AI systems themselves, their rights in relation to said systems and their possibilities of redress, complaint etc. - AI should provide an opportunity for workers to apply their skills and competences to the fullest while at the same time remain owner of the production process. This includes the principle of transparency in the use of AI systems in HR, like the hiring of employees or the performance assessment of staff. It is important to safeguard the rights and freedoms of employees in line with non-discrimination rules as regards the processing of workers' data.- UNI Europa welcomes 5.1 – 5.4. We support that these examples raise real-life concerns of the adverse consequences of AI systems. - In 5.2. UNI Europa urges the group to expand on the issue of the human's right to know they are interacting with an AI identify. This could be done through a "labelling" system. For example, online bots should be labelled as such. Users should be made aware of the use of bots and AI in customer call-centre or help desks etc. - We would welcome that the employer-employee, employer-worker relation is explicitly mentioned in 5.3 as an example of power asymmetry. - Taking into account the power asymmetry in employer-worker relations, a separate point 3.6 on "workers' rights" should be added, which should contain the following points: "decent work by design", equal negotiation processes in the sense codetermination rights, informational self-determination of employees, non-discrimination principle and freedom of association including the right to strike. This is needed in order to secure worker's rights to co-decide on aims and application of AI systems, and create a legal framework.- Concerning the long-term risks and concerns we welcome that these should be considered. This could become an integral part of the accountability and audibility demands – i.e. that developers, users deployers etc need to reflect on the	- In order to achieve "trustworthy AI", we need to establish public, independent and autonomous organisations that can control and audit (labour) algorithms (e.g. to identify underlying biases and the objectivity of data sets that train algorithms). Likewise, the implementation of the ethical guidelines on AI must be monitored. A European observatory focusing on the ethics in AI systems could play the role of an independent watchdog, including in business.- We would like the advice „to always keep record of the data that is fed to the AI systems" from the heading of data governance included under Accountability. For workers, it is paramount that the datasets used to evaluate performance, or in hiring or firing processes is transparent and can be accounted for. - The explanation of the principle of autonomy covers the question of AI at work only in a footnote, whereas this is an important issue that should be given a more prominent place. We would like to highlight the right of workers to individually and collectively opt out or withdraw from the use of AI systems (or a decision chosen by an AI system) if they undermine the workers' autonomy, decision making competence or disrespect fundamental rights and ethical principles. We recommend the inclusion of a special chapter that provides for ethical guidelines on AI in the work environment to address these issues more in detail.- UNI Europa welcomes that the HLEG on AI acknowledges the importance of social dialogue to realise trustworthy AI. We would like to add that the involvement of social partners, and in particular employee representatives, should not only take place regarding the general public debate on AI. Social partners should be involved in the establishment of codes of conducts, of standardisation schemes, development of training and in the proposed accountability governance. Employee participation and inclusion should take place early in the design, development and deployment of new technologies including AI and robotics. It is essential and important not only to inform and consult workers representatives in the work place or at branch level, but to enhance their co-determination rights and ensure their right to co-decide on the aims, reasons and implementation of AI at the workplace. - Social partners at all levels should be involved in the implementation at company, industry, national or international level, including through collective agreements setting standards. In this context, it is required to describe the negotiation processes, e.g. central control structures for sector-specific solutions (cf. ‚AI Now' Report 2018): „Governments need to regulate AI by expanding the powers of sector-specific agencies to oversee, audit, and monitor these technologies by domain." (https://ainowinstitute.org/AI_Now_2018_Report.pdf)- Regarding the principles of accountability and transparency, we need to establish mechanisms for the protection of whistle-blowers who disclose the risks of AI systems or the non-respect of ethical principles – especially in the case of employees in companies that develop such systems. Internal reporting of risks and violations should be supported and rules in place to ensure follow up. - Organisations and companies should pay attention to potential biases encoded in the system	- We welcome that processes shall be examined in order "to allow a human control, if needed" (assessment list – governing AI autonomy). In so doing it should not be a question "to keep a human in the loop". We need clearly defined measures, that empower people to exert this control in all processes – regarding resources (technical equipment etc), organisational needs (time, liability, etc.) and qualification.- We suggest extending the list on the assessment of use cases (p.28) and add the question of processes, in order to use AI to ensure decent work (development and impact assessment).	- UNI Europa welcomes the call for Accountability Governance on page 21. The establishment of Data/AI Governance Councils in companies will indeed strengthen the accountability of AI systems and will address a weakness in the GDPR. The Council should consist of shop stewards and management and be responsible for holding management accountable and transparent to the use of AI and data. Whistleblowers should be able to address concerns to the Council and mandate the council to investigate on reported issues. - We welcome the process of developing guidelines for a trustworthy AI made in Europe, which encompasses corresponding "guidelines made in Europe", but would like to raise the question why non-European companies such as Google were granted full membership and full participatory right in the High- Level Expert Group. The status of associate expert would be more appropriate.- UNI Europa also supports the position of the ETUC regarding this consultation.

				is therefore important to integrate the aspect of the quality of jobs, decent work and social progress into the ethical approach in order to create a balance with the purely economic objectives underlying the creation and use of AI and robotics. - The Human-centric approach (HCD) not only presupposes information, transparency, participation and traceability, but also requires specific negotiation processes regarding decision-making in view of the aims and implementation of AI-systems at a very early stage for stakeholders such as employees and their codetermination bodies.	development/changing nature of the adopted AI as well as engage in predictions/forecasts of its future development scope and the consequences (positive and negative) hereof). - AI's influence does not only affect the world of work, but also democracy and society as a whole. We welcome that the draft refers to this point in Chapter I, paragraph 5.3., by stating that AI is not to be implemented in order to enable "citizen scoring" by a state/government. But this should also apply to private businesses. Neither states nor companies should be allowed nor have the possibility, to create human profiles such as "moral personality" or "ethical integrity". We reject the proposed opt-out-function and even possible "opt-in"-functions are not to be designed in a way that they conflict with fundamental human rights and possibly lead to the waiving of services that are useful for a person. AI-based services, that are important for work and life, must be designed in such a way that they do not require the collection of data which could be useable for human profiling.- Creating big data-bases always includes the risk of hackability as well as intentional and unintentional data-leaks. The guiding principle of "data-sovereignty" needs data-security in order to be viable. This implies explicitly not surveying data in areas that are of highly explosive nature for people in e.g. political, private or work-related areas. Fundamental rights as informational self-determination, the freedom of association and freedom of speech are not to be put at risk by creating such data-bases.	development, training data and model performance – especially those that my affect the most vulnerable. They could also establish an internal ethical review process to democratise the decision-making process- Companies should not only increase transparency regarding the design and development of AI systems, but also in organisational chains of responsibility.	
European Group on Ethics in Science and New Technologies	EGE					In the context of the consultation process, please find the written comments arising from the deliberations of the European Group on Ethics in Science and New Technologies (EGE) in the form of an open letter via the link below: http://ec.europa.eu/research/ege
Thiébaut	Weber	European Trade Union Confederation				Artificial intelligence (AI) – or more accurately: automated decision-making - is already in use all over the EU, even if it is invisible. Often its workings are deliberately opaque in order to protect – open and hidden – corporate interests, for instance in 'social scoring', credit lines, social bots, nudging. AI is not just about technology or software programs, but societal choices are incorporated in this automated decision-making. A debate about discrimination, equality, social justice, participation in relation to AI is needed. It should be clear that AI should not discriminate, it should strengthen equality, enhance social justice and participation. Such a comprehensive approach can't be limited to ethics. The debate needs contributions from sociology, philosophy, political science, economics and data experts. The focus of the discussion must be on the politically relevant questions – at national and at EU-level. What is needed: sustainable AI made in Europe - ecological, fair, inclusive. The ETUC welcomes the approach to connect AI with European values and principles. This is a first step in the right direction, but more steps are needed. New technology, and in particular Artificial Intelligence , must be shaped in way to avoid a threat to

democracy and functioning markets. First and foremost, it has to be determined which of the challenges posed by AI can be addressed by enforceable rules and laws and which can be left to unenforceable ethic codes, guidelines, self-regulation or voluntary self-commitments.
In modern democracies it must be a principle that its cornerstones, the principles of democracy, the rule of law and human rights, must from the outset by design be incorporated in AI.
Citizens and workers, in particular workers' representatives in companies and public administration must be empowered to understand the new challenges ahead and be enabled to find appropriate answers. The GDPR was a first step in the right direction, but more regulation is clearly needed (for self-driving cars, face recognition, drones etc.) . The Commission should play a role to launch such a holistic debate involving a wide range of stakeholders and contribute to close the gap between Member States.
The focus of ETUC lies in the world of work, in particular the future of work. AI needs to be embedded in decent work. AI is ambiguous and needs to be shaped, it can be used to cement power asymmetries or to dismantle them.
It is in the interest of workers that information, consultation and board-level participation rights as well as collective bargaining are respected and fully applicable. A general information of stakeholders is clearly insufficient. The rights to information, consultation and board-level representation must cover the area of AI. A technological and social impact assessment is necessary as well as participative research to follow the design, application and implementation of AI and its economic and social consequences. It is of utmost importance that enforceable regulation creates an appropriate framework for AI in Europe.
The ETUC subscribes to a 'human-in-command'-approach to AI so that final decisions are taken by human beings and not algorithms. AI as digitalization in general has the potential to liberate work from dangerous, monotonous and repetitive tasks, in the same time allowing surveillance and control through handhelds, sensors, wearables in a totally new dimension. In order to harvest the potential and to minimise risks, it is necessary that trade unions and workers' representatives in general, and in company boardrooms in particular, regularly scrutinise and closely monitor the introduction of new technologies and AI. In particular it is important to ensure that AI fits with the targets of EU climate, energy and environment policies.
AI cannot work in a lawless zone where chatbots are not identifiable, can contribute to hate speech, influence democratic elections and undermine democracy itself. In view of the upcoming European elections, but also in democratic discourse generally, it is important to know whether one's counterpart is a human or a machine, which is not the case currently.
The rules for AI are not yet in place and it is important to take the necessary steps.
The respect for human rights, for workers' rights, for humans' moral and physical integrity, and the cohesiveness of our societies are fundamental goals, which cannot, and should not, be left to the free

appreciation of businesses regarding their marketing or communication strategy. The reaction of the EU to the very real threats posed by AI to the achievement of these goals cannot restrict itself to indicative guidelines, with no external scrutiny and no sanction in case of non-compliance.
The ETUC thus demands strong, enforceable, regulation of AI, based on legislation. EU-wide legislation has the advantage of preventing downward regulatory competition among Member States.
The legislation should prescribe procedural steps and institutions within organisations to ensure the trustworthiness of AI applications (under the model set by the GDPR), which can be verified by any layperson, and should limit to a maximum "ethics panels" or "boards", often self-serving, which do not provide sufficient predictability of their decisions and/or are vulnerable to conflicts of interest.
This legislation should bear upon the following aspects, in addition to those already described in the document:
* human workers must be able to take decisions different from the "recommendation" made by the AI system, and yet not be sanctioned for having done so when this decision proves to be wrong;
* human workers must be able to test, experiment and innovate, even against the "recommendation" made by the AI system, and yet not be sanctioned for having done so when the test / experiment / innovation fails;
* AI systems must be sufficiently reliable and their behaviour must be reproducible enough to ensure safety of material systems (specifically: of machines in a working environment), and particularly of "safety critical" systems where failure is known to cause deaths in large numbers (e.g. civil aviation, rail equipment, chemical plants, civil nuclear power);
* AI systems must only be deployed in safety-critical applications after the level of explicability of the decisions, and the capacity to trace back an accident or incident to its cause, are sufficient for this cause to be treated, and for the safety of the application to improve over time; workers must be trained to deal with AI in particular to apply the emergency brake where necessary;
* robots (aka "chatbots") must be identified and visibly marked in all on-line debates and discussions, so as not to be mistaken with genuine human opinions, or even be prohibited from taking part in some on-line discussions (e.g. on political, social or moral issues, in particular during election campaigns);
* the added value created by AI must be distributed fairly in society and economy, specifically by making sure that the access to the data that teaches AI systems is broadly distributed among all economic players under Fair, Reasonable and Non-Discriminatory (FRAND) legal and economic conditions, and cannot be captured by digital monopolists. The requirement of "distributional fairness" must be added to the list of "Requirements of trustworthy AI" given in §II.1.

Brian	Williamson	Communnications Chambers. The views in this submission are however my own and do not represent a corporate opinion.				I raise three broad questions in relation to the proposed approach:1. It is not clear that AI needs an "ethical purpose". Electricity doesn't have an ethical purpose, nor does computing. The market (what consumers want weighed against what it costs) and public policy (internalisation of externality etc) work to provide an ethical as well as economic framework for decisions. Is something different justified, if so, why is AI the relevant class of things to which it should apply? AI appears either too broad or too narrow to justify a specific approach.2. It is not clear that it is sensible or desirable to promote trust in a broad class of technology. Consumers need ways of discerning what/who is trustworthy - which may be an application or provider; but it seems unwise to promote trust in AI (or the internet or computers...). Consumers do not need to trust a technology as a whole in order to adopt it.3 It is not obvious why all AI should be technically robust to any standard. AI controlling a nuclear power-station, yes; helping me put events in my diary, surely that should just be left to consumers and the market. The following, published prior to the HLG report, touches on these points in greater detail: https://www.researchgate.net/publication/329587215_In_search_of_the_%27Good_AI_society%27
Christopher	PAINTER	meme-machines.com			I refer to the call for comments at the foot of page 22. "We invite stakeholders partaking in the consultation of the Draft Guidelines to share their thoughts on additional technical or non-technical methods that can be considered in order to address the requirements of Trustworthy AI." and comment on the section "Stakeholder and social dialogue":- In order to enable the citizen you must understand what she is talking about, rather than her answers to questions you select. The citizen expresses opinions via social media and expects to be heard; when that doesn't happen cohesion dissolves. I write from the UK. That presents a challenge for mainstream text analytics which is derived by training against examples, and is therefore obliged to put texts into pre-set boxes. That will not identify what people are talking about. There is however a solution available that shows the value of explainable artificial intelligence. It is possible to reduce any set of texts to the key word patterns that occur too frequently for chance. This does exposes what people are actually talking about, in their own words. It turns out so far that it works on any language for which a list of common stopwords can be identified. For the EU this could be crucial. I've posted more details in the forum here https://ec.europa.eu/futurium/en/european-ai-alliance/hleg-input-request-appropriate-mechanisms-measure-broad-societal-impact-ai and a reviewed paper about the technique is	

				available here http://www.rcs.cic.ipn.mx/2016_110/A%20Lookup-Free%20Approach%20to%20Knowledge%20Extraction%20from%20News%20Feeds.pdf			
Anonymous	Anonymous	Anonymous	Quote: "While the Guidelines' scope covers AI applications in general, it should be borne in mind that different situations raise different challenges."Comment: The same system could be used in different application contexts, and thus be subject to completely different norms and regulations, and give rise to completely different risks. therefore it is recommended to separate the AI system and its use, and consider the impact of its use separately for impacts. This is the "AI is a tool" school of thought that is similar to "robots are a tool"Quote: "those dealing with AI"Comment: Who is this - Users? Designers? Developers? Regulators	4. Ethical Principles in the Context of AI and Correlating ValuesThe Principle of Non maleficence: "Do no Harm"How can you be sure you are doing no harm? Are there ever cases where "do good" and "do no harm" conflict because they affect different sub-populations?Quote: "Lastly, the principle of justice also commands those developing or implementing AI to be held to high standards of accountability"Comment: ... and deploying and using AI. Accountability depends on the situation of usage. You can use a knife to cut fruit or to kill someone, and this is not the responsibility of the designer or manufacturer of the knife, but the users.5. Critical concerns raised by AIQuote: "The following non-exhaustive list of critical concerns might therefore be shortened, edited, or updated in the future."Comment: This could be any Any AI-controlled powerful machine, e.g. self-driving car or industrial laserQuote: "This involves an ethical obligation to develop entirely new and practical means by which citizens can give verified consent to being automatically identified by AI or equivalent technologies."Comment: Completely unclear as to what the implications of giving or refusing consent are. Im not sure if this is yet known.	1. Requirements of Trustworthy AISection 1: Accountability - General Comment: Missing from this section is WHO is accountable. I expected to see it in the human oversight section but it is not there. I would argue that who is accountable is in part dependent on the application case, and what is done with the AI system. I think the key issue is the human impact of the AI system's actions that is the thing that needs to be accounted for. This is clearly dependent on the function of the AI system itself, but also depends on how it is used and for what purpose, and these are choices of the user.Section 2: Data GovernanceGeneral Comment: Provenance of the training set is important for transparency - if the training set is known, it can be analysed post-facto for biasesQuote: "symmetric behaviour over known issues"Comment: OK as far as it goes, but how do you define the symmetric behaviour? Will one person's idea of even-handed treatment of the issues in the training set be bias in another person's view?Quote: "it is therefore advisable to always keep record of the data that is fed to the AI systems."Comment: The data the system uses during operation is also relevant. If a self learning the system begins to show signs of misbehaviour, how can this be undone? Do you roll back the learning to the point of malicious data? If so you will need to know which data it learnt from in operation.Quote: "To trust the data gathering process, it must be ensured that such data will not be used against the individuals who provided the data. Instead, the findings of bias should be used to look forward and lead to better processes and instructions – improving our decisions making and strengthening our institutions." Comment: No idea how the second sentence follows from the firstSection 4: Governance of AI Autonomy (Human oversight)Quote: "It must be ensured that AI systems continue to behave as intended when feedback signals become sparser."Comment: How can you ensure this? How can you define "intended behaviour" when an AI system self-adapts?Quote: "This also includes the predicament that a user of an AI system, particularly in a work or decision-making environment, is allowed to deviate from a path or decision chosen or recommended by the AI system."Comment: The user must be accountable for their actions. If an AI system suggests one course of action and the user follows another, the user is solely responsible for the consequences. The more	General comment: detection of failure modes is important. Some may be straightforward, others may be very subtle. There clearly needs to be means to detect when AI systems fail.	Overall the document is clearly comprehensive and authoritative. A couple of inter-related themes based on the specific comments have come out:Self-adapting systems should be considered in more detail, especially in terms of their regulation, reproducibility, reliability, accountability and control. Much of the document could apply to any automated system, but self-adapting systems are one of the aspects of automation specific to AI. It is difficult to certify or guarantee the reliability of self-adapting systems because they may become unpredictable as a result of their learning and adaptation. A self-adapting system is by definition non-deterministic and this means that it can behave in ways the designer never imagined. How can designing in controls to ensure ethical behaviour be respected by a self-adapting system? A possible framework for addressing this challenge is to determine a set of basic normative constraints that allow AI systems to self-adapt and improve but remain within the boundary constraints of acceptable behaviour. Who is responsible for a self-adapting system? The designer may have been fully compliant with all ethical and regulatory constraints at the point of design, but when the system adapts as part of its operation, it may go outside those constraints, and can the designer be held accountable then for the changes to the system caused by its experiences in is operational environment? Sort of the "nature vs nurture" debate. Application contexts determine many aspects of the responsibility and ethical issues that are applicable. The same system could be used in different application contexts, and thus be subject to completely different norms and regulations, and give rise to completely different risks. therefore it is recommended to separate the AI system and its use, and consider the impact of its use separately for impacts. This corresponds to the "AI is a tool" school of thought. Aspects could include:1) Who is responsible. Accountability depends on the situation of usage. You can use a knife to cut fruit or to kill someone, and this is not the responsibility of the designer or manufacturer of the knife, but the users.2) What the human impacts are. A pattern recognition system deployed in a self-driving car has completely different impacts to the same basic technology monitoring the goal-line of a football field.3) What specific regulations are applicable (there may be general ones that apply to all, but there may

interesting question regarding possibly diminished responsibility is if the user follows the AI's recommendations and as a result causes harm.Section 8: RobustnessQuote: "Currently there is an increased awareness within the AI research community that reproducibility is a critical requirement in the field."Comment: Reproducibility is further complicated by self-learning systems. Once a system changes itself through self-learning, it is different and cannot be guaranteed to produce the same results as it did in its previous state. Does this mean we need to record all states of a self-learning system?Section 9: SafetyQuote: "Processes to clarify and assess potential risks associated with the use of AI products and services should be put in place."Comment: As well as risk detection and mitigation strategiesQuote: "Moreover, formal mechanisms are needed to measure and guide the adaptability of AI systems."Comment: This is an important point, as it concerns the constraint of potentially unpredictable systems. This needs further investigation.2. Technical and Non-Technical Methods to achieve Trustworthy AIQuote: "continuously evolving and acting in a dynamic environment"Comment: Does this mean that the development of AI is evolving or that an AI system self-adapts to its environment? Both are valid, but it is not clear here.Figure 3: Comment: How does the self-adapting or self-learning aspect of AI affect this figure, particularly the evaluation & justification cycle?2.1 Tech MethodsQuote: "Importantly, evaluating the requirements and implementing the methods should occur on an on-going basis." Comment: Not clear what this means.Quote: "Central therein is the idea that compliance with law as well as with ethical values can be implemented, at least to a certain extent, into the design of the AI system itself."Comment: Easy to say, very difficult to do, I would imagine. This ties in with my previous comment about the need for normative constraints on AI systems that allow them to self-develop, learn and adapt whilst remaining within acceptable boundaries.Quote: "The requirements for Trustworthy AI need to be "translated" into procedures and/or constraints on procedures, which should be anchored in an intelligent system's architecture."Comment: ... that guide and constrain its behaviour at runtime. Again, this is easy to say but hard to do. How do you decide which constraints are relevant? How do you detect transgression? Are there basic principles or constraints that all AI should respect?Quote: "For such architecture to be adapted to ensure Trustworthy AI, ethical goals and requirements should be integrated at "sense"-level in a way that plans can be formulated that observe and ensure adherence to those principles."Comment: Surely ethical goals and requirements need to be integrated at "plan" level. "Sense" is about understanding the environment. "Plan" is about making the decisions, and here the normative constraints need to be integrated.Quote: "Testing and validation of the system should thus occur as early as possible and be iterative, ensuring the system behaves as intended throughout its entire life cycle and especially after deployment."Comment: Are you saying that the testing should be ongoing throughout the whole lifecycle of the

be specific regulations applicable to specific situations)4) What the possible failure or transgression modes are, and how they are rectified if they occur.A shameless plug: I did a consultation with experts on the specific topic of Responsible AI in mid-2018, which may be relevant for this study. The result is a report, available at: Taylor, Steve, Pickering, Brian, Boniface, Michael, Anderson, Michael, Danks, David, Følstad, Asbjørn, … Woollard, Fiona. (2018, July 2). Responsible AI – Key Themes, Concerns & Recommendations for European Research and Innovation (Version 1.0). Zenodo. http://doi.org/10.5281/zenodo.1303253Finally, I hope these comments are useful.

					system? If so this is going towards an independent monitoring system alluded to previously when disucssing checking the system against behavioural limits at runtime.Quote: "Moreover, it should be performed by an as diverse a group of people as possible"Comment: This pre-supposes that people will be doing the testing. What about automated testing? Not clear why the testing group of people should be diverse. I am not saying that this is incorrect, simply that the statement needs justification.2. Non-Technical MethodsQuote: "Many regulations already exist today that increase AI's Trustworthiness, such as safety legislation or liability frameworks."Comment: There are some generic regulations, but other regulations will be determined by the application case and the use the AI is put to. The use may not be known by the designer, especially in the event of what may be called "misuse" from one perspective or "redeployment" from another.		
Anonymous	Anonymous	Anonymous				Assessment of the common sense level and of the critical thinking level of AI system under evaluation Please consider adding the assessment of the common sense level and of the critical thinking level of AI system under consideration/evaluation. This could permit to evaluate better how some system is ready to take more or less autonomous actions/decisions.	
Julius	Kravjar			3. Fundamental Rights of Human Beings, 4. Ethical Principles in the Context of AI and Correlating Values Fundamental rights and principles described. Why values are not described?	2. Data Governance At this stage there is no mention about data formats. Definition of data is mising. Is it OK?		
Anonymous	Anonymous	Anonymous	Given the complexity of the subject (not only AI but ethics in general) and its far reaching implications on life for all of us, setting a consultation period from 18 December 2018 to 1 February 2019, is far too short . We therefore strongly call on the European Commission to continue the dialogue with a varied set of stakeholders and involve in particular representatives of SMEs in the further political and legislative process on European level. Only in this way a robust and sustainable framework for AI	The draft guidelines stay very ambiguous about a possibility of a future legal form of its ethical principles and guidelines. We believe that a set of ethical principles and guidelines without any actual level of commitment but as a sole factually non-binding guide to the developers, trainers and users of AI, will be without any use. Besides, any ex ante defined principles run the risk of being out of sync with the actual circumstances and challenges in the medium and long term.		A key problematic point of the draft guidelines is the fact that the numerous ethical principles are ultimately all thwarted by the imperative that the criteria for assessing Artificial Intelligence should be comprehensible and the consent of the person concerned (combined with a massive asymmetry of knowledge) given, regardless of whether an ethical minimum standard has actually been achieved or not.	The problematic side effects of AI are currently not explored adequately and simply unknown. For this reason we plead for a continuous discourse on ethical criteria and – as mentioned above - for an inclusion of stakeholders such as the crafts who will be not only be affected by automation and digitalisation to a very high degree in future but already showcase some notable AI-based applications in their business models.

			can be achieved. It has been agreed that the objective of the guidelines is to develop a specific set of ethical rules for "trustworthy AI made in Europe". If this is the case, it is irritating to see that the expert group (and reserve list) has included US internet companies (Google, Amazon) as their members. Undoubtedly it is to be feared that the concerned companies could influence the discussion in a way to level down any sort of "ethics barriers" for their offer on the European AI market or to keep them as small as possible from the very outset. In addition to representatives of science, the group is mostly dominated by companies and associations whose interests lie primarily in the rapid and broad market penetration with AI solutions. Representatives belonging to the "affected" group rather than to those of the developers and providers of AI solutions don't appear on the list , although the ethical framework speaks explicitly about the protection and benefit of and for users.					
Ana-Maria	LLORENTE	Insurance Europe	Insurance Europe acknowledges the importance of developing, designing and deploying a human centric and ethical AI and, therefore, we welcome the AI HLEG efforts in setting up a draft guidelines for Trustworthy AI. Moreover, we welcome the opportunity to react to the draft guidelines. However, we regret the consultation period, which, even if extended by two weeks remain much too-short, and we urge the European Commission and its expert groups to ensure that a reasonable consultation period (for example 8 weeks) is set for the next steps. Importantly, Insurance Europe would kindly request the AI HLEG to hold a second consultation on the final draft guidelines, which is expected to be published in March. This would allow relevant stakeholders to comment on the three use cases related to the insurance industry that the expert group plans to include into the final guidelines (reference in page 27 of the draft guidelines). While Insurance Europe is fully aligned with the proposed ethical approach grounded on fundamental rights, we would like to put forward the following preliminary comments:- We appreciate that the HLEG' s recognises that there is currently no legal vacuum in Europe given the existing many regulations that apply to AI and its use (page 2). The insurance industry is a highly regulated and supervised sector at national and European level and, consequently, we believe that a responsible and Trustworthy AI is already ensured in our industry. Moreover, the insurance sector is very experienced in using data and new technologies in a responsible and secure manner. For further details on the insurance regulatory framework, please consult our Q&As paper on the use of big data analytics in insurance (see link in https://www.insuranceeurope.eu/qas-use-big-data-insurance).Therefore, we recommend that the existing regulatory framework is duly considered when developing the use cases related to insurance premiums.- Insurance Europe agrees with the HLEG's assessment of the potential benefits of AI for society and human well-being. However, to exploit these benefits, it is crucial that any framework regulating and/or providing guidance for AI	Insurance Europe believes that the premise for a responsible use of AI and technology is not only full compliance with fundamental rights as recognised in the European Charter of Fundamental Rights, the European Convention on Human Rights and any Human Rights Convention signed within the framework of the United Nations, but also fundamental rights assimilation across industry corporate governance and within any department involved in the design, development and deployment of AI.However, in the development of a sound Trustworthy AI, it may be necessary to consider some challenging trade-offs. In this regard, Insurance Europe would like to draw the expert group's attention to an article published by the International Association of Privacy Professionals (IAPP), which explains how algorithms can reduce discrimination through the processing of "proper data" (see link in https://iapp.org/news/a/algorithms-can-reduce-discrimination-but-only-with-proper-data/). It is the author's view that making biases transparent by the processing of sensitive data can be key in eliminating biases and thus discrimination of vulnerable groups. The expert group could consider whether discrimination and inequality could be better addressed and avoided, if their mapping were possible through the assessment of "proper data", including the categories of data identified in the GDPR as sensitive data (e.g. race, gender or religion).However, if this approach were to be considered, it should always be combined with a reasonable analysis of the results. In this regard, sensitive categories of data can sometimes be correlated with other categories of data that may have an impact on the decision. For example, there are some professions that are highly dominated by one gender, while at the same time a person's profession may be an essential risk factor, e.g. in the case of disability insurance, because of its potential impact on the individual's health (e.g. miners). This correlation may seem to show a gender effect, however a reasonable analysis would show as not being unfair discrimination but based on risk analysis, and therefore it should not be automatically eliminated.Regarding the principle of	Insurance Europe welcomes the expert group's selection of the ten requirements for a Trustworthy AI, as we understand these requirements can play a key role in providing guidance for a responsible use of AI. We would like to put forward our thoughts on some of the requirements from an insurance-specific point of view and to share with the expert group a few general comments:- Governance of AI Autonomy (human oversight): The draft guidelines rightfully point out in page 15 that "the greater degree of autonomy that is given to an AI system, the more extensive testing and stricter governance is required". Moreover, they provide in footnote 24 the different layers that AI autonomy can present. Insurance Europe supports the footnote, and suggests that it should be part of the main text in the section for requirement 4 "governance of AI autonomy". Moreover, and considering the different levels of AI autonomy, we believe that the expert group should further emphasise a "risk-based approach" throughout the draft guidelines. In this regard, the amount of measures that an organisation should put into place to prevent any potential risks should depend on the level of autonomy of the AI device. For example, the decision process reevaluating the continued operation of a life-support-machine has significantly different consequences than an advertisement shown on a social media timeline (e.g. add shown in the Facebook application). Consequently, the AI governance processes behind these two examples would necessarily be different.- Non-discrimination: Insurance Europe would like to highlight that the basic principle of insurance is the accurate assessment of risk and would encourage the expert group to distinguish between fair risk assessment and unfair discrimination. Industries, including insurance, should be able to use machine learning methods to perform dynamic pricing provided that they have data governance processes in place to ensure that factors which are legally prohibited (e.g. discriminatory factors) are removed from the decision-making process. - Respect for and enhancement of human autonomy: The draft guidelines state in page 16 that "AI products	Insurance Europe's thoughts on the case studies are reflected in the section below (general comments).	Insurance Europe acknowledges the importance of developing, designing and deploying a human centric and ethical AI and, therefore, we welcome the AI HLEG efforts in setting up a draft guidelines for Trustworthy AI. Moreover, we welcome the opportunity to react to the draft guidelines. However, we regret the consultation period, which, even extended by two weeks remain much too-short, and we urge the European Commission and its expert groups to ensure a reasonable consultation period (for example 8 weeks) is set for the next steps. Importantly, Insurance Europe would kindly request the AI HLEG to hold a second consultation on the final draft, which is expected to be published in March. This would allow us to provide our views on the three use cases related to the insurance industry (out a total of four cases) that the expert group plans to include into the final guidelines. These use cases are (page 27 in the present draft guidelines): autonomous driving/moving, insurance premiums and healthcare diagnosis and treatment. Notwithstanding the above, and as requested by the expert group on page 27, we would like to put forward our preliminary thoughts on the assessment of the use case related to insurance premiums. It is crucial that the expert group considers the specific nature of insurance and adequately interprets and adapts the ten requirements for Trustworthy AI to the insurance business model and insurance products. Therefore, the interpretation of fairness should closely consider the principle of risk-based pricing, which is fundamental for effective insurance markets, reliable insurance cover for customers, and insurers' financial stability. For further details on how the insurance business model and the insurance principle of risk sharing work, please refer to Insurance Europe's Q&As paper on the use of big data in insurance (see link in https://www.insuranceeurope.eu/qas-use-big-data-insurance).

design, development and deployment not only aims to prevent potential prejudices to fundamental rights and their safeguards but also provides a future-proof and innovation-friendly framework. Any such framework should not hamper effective competition and companies' efforts to serve their customers fairly. We also call on the HLEG to integrate the principles of technology neutrality and proportionality in its future guidelines for Trustworthy AI in order to guarantee the effectiveness of these principles while protecting fundamental rights and promote Europe as a global technological competitor. In other words, the ten requirements in the draft guidelines and which Insurance Europe supports should be proportionately applied, irrespective of the technology used. These principles are also outlined in the Commission's FinTech action plan.- Overall, the draft guidelines mainly focus on the perspective of individuals/citizens and how they could be affected by AI. While Insurance Europe agrees that this should be the primary focus, it should be noted that many companies in a business to business context (B2B) do not have direct contact with individuals/ citizens. Nevertheless, Trustworthy AI remains extremely important in a B2B context. What changes is the focus on the different requirements. For example, in a B2B environment, accuracy of algorithms, robustness and transparency become more relevant than the ethics related to individuals. Insurance Europe recommends that the expert group acknowledges the B2B context in the draft guidelines as this is currently missing. The expert group should also consider how the focus of the different requirements should change as regards secondary applications of AI. (e.g. the use of AI to expedite the payment of insurance claims) since these are not currently considered. - Insurance Europe supports the proposed framework for a Trustworthy AI, as illustrated in figure 1 in page 4. However, we believe that the proposed structure lacks the adequate level of connection between the different layers. In other words, the ethical principles described in Chapter I are not always sufficiently linked to the ten requirements for Trustworthy AI in Chapter II, and finally the connection between these two layers is not linked enough to the final layer concerning the technical and non-technical methods that should enable a Trustworthy AI. For example, the connection between the principle of justice (be fair) and its link with the requirements for Trustworthy AI is unclear. We presume that this principle is related to the requirement of non-discrimination and the requirement for accountability. However, the explanation that provides the connection between the different layers is not sufficiently exhaustive.Therefore, Insurance Europe recommends the expert group to revise the connections between the different layers and where needed to provide further clarifications allowing a clearer interpretation.- The guidelines, can be a helpful instrument in centralizing a human-centric approach, when considered together with the existing laws that protect citizens from potential harm caused by the use of technology (e.g. GDPR). Insurance Europe would welcome further information on how the guidelines will be enforced and its principles upheld among businesses

explicability ("operate transparently") the draft guidelines include the following statement in page 10: "'Individuals and groups may request evidence of the baseline parameters and instructions given as inputs for AI decision making (the discovery or prediction sought by an AI system or the factors involved in the discovery or prediction made) by the organizations and developers of an AI system, the technology implementers, or another party in the supply chain". Insurance Europe would like to highlight that while providing meaningful information is part of the transparency obligations of data controllers (Articles 13 and 14 of the GDPR), and a prerequisite for obtaining valid consent (Article 4(11) of the GDPR) as well as a well-established right of the data subject , this should not adversely affect the rights of other parties, including protection of trade secrets, intellectual property rights or for example, the processes developed by an insurer to detect fraud or the company's know-how. Insurance Europe believes that the right balance between citizens' rights and companies' rights should be found when assessing what exactly the provision of "meaningful information" entails. Regarding the section on "normative & mass citizen scoring without consent in deviation of Fundamental Rights" the draft guideline state in page 12 that "whenever citizen scoring is applied in a limited social domain, a fully transparent procedure should be available to citizens, providing them with information on the process, purpose and methodology of the scoring, and ideally providing them with the possibility to opt-out of the scoring mechanism. This is particularly important in situations where an asymmetry of power exists between the parties. Developers and deployers should therefore ensure such opt-out option in the technology's design, and make the necessary resources available for this purpose." From an insurance perspective, it is very relevant that the risk ratings of individuals are not considered a form of scoring as described in the draft guidelines. This is because the basic principle behind the insurance business model is the accurate assessment of risk and risk differentiation, which is later reflected in the price (further details in Insurance Europe's big data Q&As). It would be challenging to achieve full transparency as proposed by the draft guidelines in the insurance sector, due to commercial sensitivity and conflicts with Intellectual Property laws. Equally, providing opt-outs could lead to situations of moral hazard where individuals with higher than average risk profiles would opt out of any risk-rating mechanism. Similarly, this principle would limit the use of credit scores for loans and other financial services that have been long used before the development of AI capabilities. Therefore, Insurance Europe urges the expert group to carefully reconsider the section on "normative & mass citizen scoring without consent in deviation of Fundamental Rights" with the view on the possible impacts on established and respected methodologies in use in the financial sector.Finally, the expert group expresses their fear for potential longer-term concerns, which today are not yet identifiable. Insurance Europe would support the setting up of a monitoring system that could on a yearly basis identify and warn of any possible risks or dangers related to the

and services, possibly through "extreme" personalisation approaches, may steer individual choice by potentially manipulative "nudging". At the same time, people are increasingly willing and expected to delegate decisions and actions to machines (e.g. recommender systems, search engines, navigation systems, virtual coaches and personal assistants). Systems that are tasked to help the user, must provide explicit support to the user to promote her/his own preferences, and set the limits for system intervention, ensuring that the overall wellbeing of the user as explicitly defined by the user her/himself is central to system functionality." Insurance Europe understands the expert group's concerns to preserve human autonomy in an AI environment, however, we believe that this section presents an unnecessarily negative view on nudging. Insurance Europe notes that the concept of positive nudging is not considered in the draft guidelines (e.g. helping customers reach their personal goals such as exercising more, improving their health etc.). The insurance industry sees a mutual benefit in encouraging people to behave in a way that improves their health: from an individual's perspective, positive nudging helps individuals achieve their human right to health, and from the insurers' perspective, healthier individuals contribute to better risk pools. Moreover, nudging is already widely used in advertising, retail, and many other business areas – both in the online sphere as well as in traditional commerce. Therefore, the expert group's statement seems to unfairly disadvantage AI development. Insurance Europe recommends the expert group to review the paragraphs on page 16 to present a more balanced and realistic approach on nudging. Finally, Insurance Europe would like to briefly comment on the proposed technical and non-technical methods to achieve a Trustworthy AI:- General remarks: It is our view that non-technical methods should be introduced in first place in the draft guidelines and followed by the technical methods as non-technical methods are more important in achieving a Trustworthy AI. An organisation firstly needs to assimilate the ethical principles, for instance, via a code of conduct, education and awareness campaigns, to allow for the technical methods to be a success. In other words, the non-technical methods shall drive the technical methods.- Additional non-technical methods to achieve a Trustworthy AI: Insurance Europe is of the view that, although effective competition in a market economy does not have an answer to every problem, it can be of great assistance for achieving Trustworthy AI. Even today, the reputation of a company and company ratings by market intermediaries are important drivers of customer decisions. For example, a company that complies with laws, but does not act ethically, will face reputation issues and many customers will most likely choose another provider. Therefore, these tools can act as powerful non-technical methods in achieving a Trustworthy AI. However, the trade-off caused by the use of these tools is the risk of facing the effects of "misinformation". Unjust and untrue publicity could slow down the development of innovative products in Europe, due to fear of being subjected to these sorts of negative and unfair

				operating in Europe. However, Insurance Europe wonders how European Institutions are going to ensure that the guidelines are adhered to and effectively applied not only by EU based companies but especially by non-EU businesses that offer their products and services in Europe. We would very much welcome further information in this regard.	use of AI. The HLEG could have the lead on this monitoring system.	campaigns.The ideal scenario would be one, where high quality ratings and reliable information on insurers' offers are available and place companies behaving unethically in the public spotlight, while authorities can control the spread of "misinformation" and its negative effects.- Technical methods-testing and validating: Organisations should not only include security tests as described in the draft guidelines but also testing to avoid non-discrimination and bias. - Technical methods-traceability and auditability: Insurance Europe agrees that modelling techniques should allow companies to at least extract the major factors that influence decision-making processes. However, it may be too restrictive and burdensome to suggest that companies should make the causality of decision making comprehensible to a layperson.	
Joint UNICEF comments	Joint UNICEF comments	UNICEF	No specific comments to this section.	This is a great chapter, and UNICEF fully supports using the rights-based approach to AI Ethics.As part of the rights-based approach it is important to consider different groups of people (diversity aspect), especially children and youth as one specific group whose perspective is often forgotten even in the human-centric approach. The rights of children need special protection. In addition to children being a vulnerable group in our society, children often do not get their voice and viewpoints heard.Therefore we suggest integration of children and their rights in the following parts of this chapter:3. Fundamental Rights of Human Beings- We suggest adding to the list the best interest of the child, as per the Convention of the Rights of the Child. The Convention is the most ratified human rights treaty in the world, and is also ratified in every country in Europe. The best interest of the child is one of the key principles of the Convention, and states that the best interest of children must be the primary concern in making decision that may affect them. 4. Ethical Principles in the Context of AI and Correlating ValuesIn the introductory text:- It is a great suggestion to have the presence of an internal and external (ethical) expert to accompany the design, development and deployment of AI, given the potential of unknown and unintended consequences of AI. It would be importance to ensure that this person is trained of diversity, and also on specific considerations of different groups, including children.Under the listed principles and values:- The Principle of Non-maleficence: "Do no Harm": We recommend this principle to bring up also the tension related to current social media business models based on attention harvesting for	Our comments to the Requirements of Trustworthy AI:• Data Governance: Gathering and using data from and about children is especially sensitive and extreme care must be taken. See UNICEF's Tools for Business as useful resources: https://www.unicef.org/csr/ict_tools.html• Designing for all: It is important to consider children and youth also as one group. This part should also refer to the Convention on the Rights of the Child, in addition to the UN Convention on the Rights of Persons with Disabilities.• Transparency: A challenge with the ICT industry is how to balance transparency against commercial intellectual property. Companies may not want to be transparent at the risk of losing their business advantage.Our comments to Non-Technical Methods:• Accountability Governance: The suggested idea of an internal and/or external ethics panel or board is great. It would be important that the representatives of the panel/board come with various background and expertise, and that at least some of them would also understand child rights, among other issues.• Codes of Conduct: It would be great to include some examples of what good KPIs might look like.• Education and awareness to foster an ethical mind-set: It is important to provide the user and impacted group education and awareness-raising on the potential impact of AI also for children and youth so that they grow to be responsible, informed and active digital citizens, that are part of shaping the society.• Stakeholder and social dialogue: It's important to ensure that children's views are heard, and children can also participate in the stakeholder consultations, not only adults.• Diversity and inclusive design teams: It is important to	No comments.	We at UNICEF congratulate the High-level Expert Group on putting together this great draft, and we especially appreciate the rights-based approach, as well as the human-centric approach. Our main comment is that even if human-centric approach should include everyone, often in practice only adults are considered and children are forgotten, unless special attention is paid to them. Therefore, through our more detailed comments to the various sections we have tried to make sure children and youth are not forgotten but brought more to the forefront of the discussions.We at UNICEF would also like to point out that while children should be considered as one of the vulnerable groups that need special attention, children are also active members in our society and should also be empowered to be the current and next generation AI users. We did not notice any representation of child / youth voices in the expert group that has put together these guidelines. We would recommend ensuring that also children and youth voices are heard in this process.Overall, the guidance would benefit from examples of best practices. For instance, it might be worthwhile to include an example of a case where information to stakeholders (customers, employees, etc.) about the AI system's capabilities and limitation has been provided in a great manner.We would also be very interested in knowing more about the implementation strategy, monitoring mechanisms and resources related to the operationalizing of the guidelines.

			advertising vs. what is healthy, private behavior for users.- The Principle of Explicability: “Operate transparently”: The principles of technological and business model transparency are important, but not easy to get right when communicating with a non-technical audience. Please see the Allegheny Family Screening Tool as an excellent case study on this: http://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx Under the Critical concerns raised by AI:- Identification without Consent: The age of consent of children, as also per GDPR, is a very relevant point to consider here. In many cases, age verification systems do not currently work in a trustworthy manner. Also, as the document points out, consumers give consent without consideration. This is especially true in case of children, as they may not even understand what they’re giving consent to. Based on the evolving capacities, children of different ages may not have enough understanding or information to assess the long-lasting impacts of how the data they’ve given consent to may be used. - Normative & Mass Citizen Scoring without consent in deviation of Fundamental Rights: It would be important to provide people (including children) the opportunity to opt-out of the scoring mechanism. However, opting out may not always be possible due to the existing systems and practices. A case of where a school district uses a digital learning management system is a good example. In such cases there should be alternative option. In this case the opportunity to question the scoring system or request human intervention should be also considered.	consider how children’s views are also considered in this context, especially because normally the teams designing, developing, testing and maintaining the systems are adults.Additional ideas of technical and non-technical methods: Some great new ideas we’d like to put forward are "Child Rights Ratings Tool for AI-content/platforms" and "Child-friendly algorithms", which could carry some form of certification. These ideas were discussed in an AI workshop organized by UNICEF and World Economic Forum. These and other ideas can be found from here: https://www.unicef.org/innovation/stories/generation-ai	
Tuesday	Porter	TÜV NORD AG			TÜV NORD supports the general principles set out in Chapter 1. We support the holistic approach taken to draft the guidelines for Trustworthy AI, as the impact of AI is highly dependent on the environment in which it is used. Thus, based on the principles set out by the HLEG, the guidelines need to balance between the application of AI in basic research and in internal company processes, where innovation should not be hindered by red tape, and the establishment of clear rules for AI in areas in which citizens’ fundamental rights might be affected. Thus, TÜV NORD supports the notion that AI should be held to high standards of accountability, and that AI systems should be auditable (p. 10). Based on this holistic view, TÜV NORD is in favor of a risk-based approach to the establishment of Trustworthy AI in Europe, where AI that might directly impact citizens’ fundamental rights has to be held to higher standards of accountability. Following the time-tested approach the EU has taken towards ensuring the impact of technology on EU citizens, this can be achieved through the implementation of independent third-party testing and inspection in cases where the use of AI is deemed a possible risk to citizens’ fundamental rights. Said holistic approach should include many of the elements mentioned in Chapter 2 and 3 of the HLEG guidelines. Viewing the entire process from design to application, elements

							such as the integrity of the data used for AI decision-making, the quality and suitability of AI decision-making processes, the transparency of these decisions to consumers, and the IT security of AI systems all have to be considered to evaluate the possible risk level of a particular use case of an AI system. Based on such an assessment, a risk-based approach could require third-party testing and inspection for AI use cases of certain risk levels. The EU New Legislative Framework and the EU Cybersecurity Act can serve as models for an approach on how to ensure that Trustworthy AI will have no negative impact on citizens' fundamental rights. Furthermore, TÜV NORD sees the necessity to emphasize the importance of standardization, as mentioned in Chapter 2 (p.21). On a global level, ethical considerations are often not a focus of the AI standards that are currently being developed. Here, Europe has the opportunity to set standards, which could not only be applied in Europe, but by other stakeholders around the world who place importance on ethical considerations when developing and using AI systems. There are also existing standards that could be applied to AI systems, in order to adapt already existing and working standards and speed up their application in the AI realm.
Michael	Mabe	International Association of STM Publishers	STM supports the initiative for Trustworthy AI to be developed in Europe. In 2018, STM released a statement "STM Publishers Innovations Support AI and Machine Learning" at https://www.stm-assoc.org/2018_10_08_STM_Publishers_Innovations_Support_AI_and_Machine_Learning_8Oct2018docx.pdf. This statement not only supports the further development of AI but also outlines how STM publishers are moving to meet the needs of the AI era.		STM publishers are stakeholders in the development of AI, undertaking the curating and making available of scholarly information and data and continually developing new formats for which to do so, including formats that are operable with AI and AI products. As partners in AI development, STM publishers ask that the information products they bring to the development of AI, and the consequent value that they add, be appropriately recognised in the development of AI products that use this information and data.		STM welcomes the Guidelines in that they bring to the fore elements that contribute to Trustworthy AI. STM concurs that any state-of-the-art AI must pass an assessment of being trustworthy and specifically developed and adapted to the specific use cases and application domains. Trustworthy AI may involve the use of a combination of open source software tools, third party software tools, content and data for training and algorithms, as well as novel proprietary implementation. The combination of the content and data used for training and algorithms is where the power in AI lies. Quality of data fitted to the specific AI application area is key. A functioning Trustworthy AI system must be trusted by the stakeholders in that system including in that it has to respect the rights and contributions of all participating stakeholders. That implies also respect for their intellectual property rights that may subsist both in the specific implementation of an AI system and in the curated content and data that is domain-specific and required for implementation.
Thomas	Zeilinger	Evangelical Lutheran Church of Bavaria (Evang.-Luth. Kirche in Bayern)	Trust is rightly seen as a key prerequisite for a human society (p.1). However trust always has to be balanced with security whenever possible. Transparency, explainability as well as reliability should be on the side of security and not on the side of trust. Trust is a relational category: somebody trusts in s.body. As trust is stated towards a person or an institution/organisation, it cannot be AI in and out of itself who is the object of trust (AI "worthy of trust", p.1). In other words: trust needs to be concrete and specific, not generalised. Which measures do	The "human centric approach" and its underpinning in "fundamental rights" is key - as stated. The model of Beauchamp and Childress, which is applied p.8ff.was introduced and developed in the context of the health system. In that system you have a convergence of interests between the physician and the patient. However that is not the case in the field of AI, where different agents have different interests. Therefore the application of the Beauchamp/Childress model towards an	The chapter lists the key factors for realising "trustworthy" AI. It does stay vague however, what is expected from whom at which point. The ethical category of "responsibility" is not specified, as the terminology of the paper refers to AI with an amount of anthropomorphical terms which seem to equate human and technological actions. In ethical hindsight there needs to be a differentiation between personal and technological processes of perception, interpretation, and decision: responsibility cannot be applied to technological systems	The envisioned circular model seems appropriate. The chapter can be better judged as soon as the announced use cases will be presented.	The goal to develop a framework for trustworthy AI as the "north star" of a European approach towards Ethics in AI should be underlined, not only in an ethical perspective but as well seen economically as a medium-time advantage in the global race for AI. The self-understanding of the document as a starting point for discussion should give more time for discussion than just the first month of 2019!

			provide trust needs to be spelled out in the guidelines. p.2, line2: there is no "legal vacuum" in general for sure, however it seems debatable wether or not there is enough regulation applicable to AI in place in Europe! Therefore ethical considerations can and should not just have the quality of guidelines but must consider and put forward the question in which areas there is a need for laws and regulations "above" the level of guidelines!	ethics of AI may be dangerous in hiding the real (diverge of) interests behind a (ideological) curtain of presupposed - yet false - harmony! p.10: "explicability": which company would have a shared interest in that - unless mandated by law?!	aka AI, but to the human person designing, applying or evaluating technology.		
Ansgar	Schaefer	Universität Konstanz			Non-Technical Methods, pp. 22-23: I strongly suggest to amend the subsection "training and education". Not only should "managers, developers, users and employers" or "the public" be aware of and/or trained in Trustworthy AI. But Trustworthy AI should be a recommended topic of the curricula (e. g. in study progammes) in any professional training related to AI as well as trainings which educate persons who will be otherwise professionally affected or in contact with AI. Reason: Start early, not only, when somebody is already working on or responsible for AI. Thus, integrate issues of Trustworthy AI where future engineers and deciders are educated.		
Johan Christian	Amby	Danish Ministry of Industry, Business and Financial Affairs	The Danish Government agrees that the guidelines should complement and go beyond – rather than replace – compliance with fundamental rights and applicable regulations. Specifically, the Danish Government agrees that the guidelines should build upon and go beyond the GDPR as the relevant legal framework for protection of personal data. The AI/data ethical principles should be about creating incentives for businesses to go beyond the letter of the law and drive change because they see the competitive advantage in being ahead of the curve.The Danish Government finds it useful that the guidance is addressed to all relevant stakeholders that develop, deploy or use AI or other data-based technologies. However, there are specificities and additional obli-gations when using AI/data-based systems for decision-making in the public sector which do not necessarily apply to private entities.The Danish Government agrees that essentially ethical use of AI/data requires both ethical purpose and technical robustness. However, alt-hough human dignity should always outweigh profits, it should be clear, that AI may indeed be legitimately used to improve the competitiveness of a business – as long as the applicable regulations and ethical principles are observed.Finally, it should be considered when and how it is appropriate to use AI. An example could be in health care, where AI can provide a multitude of information on the patients based on already collected data on the patients. This can in theory be used to assess or indicate the patients' risks of developing specific diseases at a later stage in the patient's life. Already now, apps using AI exist where the risks of developing specific diseases can be scored in percentage. The questions are if and when this type of information firstly should be provided and secondly should presented to patients. Providing all information on all	The Danish Government agrees with the human-centric and rights-based approach of the HLEG implying that the starting point of the framework should be the fundamental rights commitments as set out in the EU Treaties and in the Charter of Fundamental rights. Outlining the fundamental rights which all EU Member States adhere to, serve as a good basis for developing a common EU approach to AI/data ethics. Although the principled considerations regarding the fundamental rights of the EU and the AI4People's project are very relevant, the HLEG is invited to consider, whether such considerations could be communicated in a more stringent manner or be included in an annex. To be of real use for businesses, public entities and researchers, the guidance should focus on providing concrete recommendations on the design and application of AI/data-based systems.	Overall, the Danish Government finds that the ten requirements for AI/data-based systems and applications are useful in terms of implementation and operationalisation of core values and ethical principles into AI/data-based systems. Some of the requirements seem to be slightly overlapping, which is why the HLEG is invited to ensure clarity and stringency with the end-user in mind. As to the "Design for all" requirement, the aim of enabling equitable access and active participation of all citizens when it comes to public services is indeed relevant. However, requiring that all AI/data-based systems including all private products and services be designed to allow all citizens to use them – regardless of their age, disability status or social status – could be detrimental to innovation and competitiveness for EU businesses. The issue of security is partly covered by the requirements on "Robust-ness" and "Safety". Given the importance of cyber security in relation to AI and other data-based technologies, the item should be elaborated.The HLEG both recommends removing bias and limiting bias given the fact that data always contain a certain bias. As the issue of non-discrimination and bias is very important, the guidelines should state stringently that although AI/data-based systems may never be unbiased, undesired bias should be consciously removed. The list of specific technical and non-technical methods to achieve trust-worthy AI is particularly important as a first step. Down the line the list could be supplemented by more technical recommendations to assist in the development of AI/data-based solutions. Although recognising that there are specific challenges related to AI in terms of explainability among other issues, the methods outlined by the HLEG apply more broadly to the use of data-based technologies includ-ing internal and external	The Danish Government finds that the assessment list contains many relevant items and questions to be considered when applying AI and other data-based technologies. It is important that the final version is drafted with the end-user in mind as a non-exhaustive check-list. To make the guidelines as relevant as possible the requirements and assessment tools should be concrete, simple and to the point.	General Comments Artificial Intelligence (AI), hold enormous potentials for optimization and innovation both in the public and private sector. However, they do not just pose the questions of how we are to gain from it or what we can achieve. They also pose a series of questions in relation to privacy, transparency, responsibility and even democracy. Ensuring that data are handled responsibly is a new challenge for businesses, for the public sector and the society as a whole in order to uphold the trust from citizens in the uptake of new technologies. At the same time AI represents a great opportunity for Europe. And if the EU succeeds in establishing a data ethical approach to AI that allows EU businesses to distinguish themselves positively from their competitors in the global marketplace, AI/data ethics could become a competitive advantage for EU businesses. Thus, the Danish Government welcomes the Draft AI Ethics Guidelines for Trustworthy AI, prepared by the High-Level Expert Group on Artifi-cial Intelligence (HLEG), and believes that the guidelines will be an im-portant first contribution towards a common EU approach on AI/data ethics. The Danish Government strongly agrees that AI/data ethics should be a key enabler of European global competitiveness and supports the approach of providing guidance rather than resorting to regulation at this stage. The Danish Government also supports the HLEG's aim of using the guidelines as a lever to influence the norms on ethical use of AI/data at global level. In line with the recent recommendations of the Danish Expert Group on Data Ethics, the purpose of the HLEG's work is to offer concrete guid-ance on the implementation and operationalization into AI/data-based systems of core values and ethical principles. Through national stake-holder consultations in Denmark it has become evident that

possible later diseases to a patient can lead to mismanagement of health care resources caused by possibly superfluous treatment but could also – which is more important in an ethical context – lead patients to worry unnecessarily about diseases that might never develop. Therefore adding an ethical point on when and how to use AI could be relevant in the Guidelines.

expert advice, organisational culture, auditabil-ity, traceability, training and education, standardization and transparency. As to diversity in teams, although it is very important to ensure diversity in the team developing AI/data-based solutions in terms of different mindsets and educational backgrounds, diversity in terms of gender, age, ethnicity is better tackled in the overall recruitment strategies of public and private entities.

businesses want to work systematically on AI/data ethics, but that they lack the nec-essary practical oriented tools and guidance. The Danish Government therefore commends the HLEG intentions of providing concrete tools that may guide the AI/data ethical efforts of public and private sector entities. The Danish Government recommends that the guidelines eventually should be accompanied by more technical and best practice-oriented recommendations on how to design AI/databased systems to ensure e.g. explainability. This will make it easier for businesses to incorporate and take actions on AI/data ethics in their work. Furthermore, the recommendations should be accompanied by concrete measures to strengthen transparency allowing consumers to make demands on data ethics and to select a data-ethical alternative when navigating between companies, websites, apps, services and products. The Danish Government thus believes that the guidelines – subject to a positive impact assessment - could enable the development of a European Data Ethics Seal by the relevant industry and standardization bodies. Such a seal could be possible to use for companies that live up to a pre-defined list of data ethical requirements e.g. following high standards for data security, not collecting unnecessary user, using algorithms that have been tested for biases etc. A European Data Ethics Seal could be a way to operationalize the idea of "ethics by design" and make it visible for the consumers which companies, products and services to trust and thus creating a market incentive for businesses to become more data ethical.The Danish Government supports that the guidelines will be accompanied by a mechanism enabling all stakeholders to formally endorse and sign up to the guidelines on a voluntary basis. However, that would require a document/a set of guidelines specifically aimed at being endorsed i.e. with concrete recommendations that may be implemented in practice in order for companies to put it into practice. Furthermore, the continuous revision of the document which is envisaged also needs to be considered. For example, a revision clause could be included in the document that is to be formally endorsed. The European Commission should also examine the potential in amend-ing Directive 2014/95/EU as regards disclosure of non-financial and diversity information by certain large undertakings and groups. The revision could e.g. include a requirement that certain large undertakings to prepare a non-financial statement containing information relating to their data ethics policies as part of their annual management reports. This would both create an incentive for companies to work actively with data ethics as a competitive parameter and to develop new data ethical solutions as well as provide a potential first-mover advantage for EU-businesses in the global market place. The guidelines should take into account the standardisation efforts in other relevant fora. ISO/IEC and CEN/CENELEC as well as the IEEE are currently working on different aspects of standardization of AI and eth-ics/trust. Furthermore, the Commission should investigate the possibili-ties for utilizing technical standards on AI Trustworthiness in the European legislation on AI and data technology in general, since

								standards have proven to be an extremely flexible way to regulate an industry.Finally, although the specific principles on development and deployment of AI are indeed very welcome, the Danish Government finds that the guidelines also should cover the broader application of all data-based technologies and not be limited to the use of AI systems. The Danish Government looks forward to engaging in the ongoing work of the HLEG on the AI Ethics Guidelines for Trustworthy AI due in March 2019 as well as in the work on the Policy and Investment Recom-mendations due in May 2019.
Francesca	Gaudino	Baker McKenzie	-		Comments by Baker McKenzie:Chapter I: Key Guidance for Ensuring Ethical Purpose:We believe it would be beneficial to rephrase Section I.2, titled "From Fundamental rights to Principles and Values," of the Draft Ethics Guidelines for Trustworthy AI ("Guidelines") (pages 5-6) in order to provide more clarity on the interrelationships between fundamental rights, principles, and values, especially with the claimed purpose of the Guidelines in mind: "In contrast to other documents dealing with ethical AI, the Guidelines hence do not aim to provide yet another list of core values and principles for AI, but rather offer guidance on the concrete implementation and operationalisation thereof into AI systems."General comments on sections discussing fundamental rightsFor coherence and clarity purposes, we believe that references to fundamental rights in the context of any AI guidelines should be directly tied to the EU Charter of Human Rights (Charter), with a view of eventually adding a "third generation" fundamental right relating to AI to the Charter, similar to the "third generation" fundamental rights that already exist, such as for data protection, guarantees on bioethics, and transparent administration. The introduction of an "AI transparency" fundamental right would, in its core, be based on the fundamental rights to dignity, freedoms and equality, and would reflect the already-existing fundamental rights of the protection of vulnerable groups, e.g. "rights of the child" and minority groups against biases, such as sexism, racism, xenophobia, homophobia, and other biases against minorities.Comments on sub-section on "The Principle of Beneficence: 'Do Good'" under Section I.4:We acknowledge the difficulty of providing an all-encompassing discussion on AI, but we propose including discussions on the following topics given their potential in materially effecting "beneficence":• The importance of developing and deploying AI that promotes the protection of the environment and its sustainability, as well as	Comments by Baker McKenzie:Chapter II: Key Guidance for Realising Trustworthy AI:Comments on Section II.1.1 on "Accountability"We believe that accountability governance may be further strengthened by designating person(s) to assume responsibility for actions taken by AI that result in harm. This would help realise Trustworthy AI. The issue of allocating liability is not easy to solve due to the complex nature of relationships between AI developers and AI users. In our view, various liability frameworks, such as vicarious liability and joint responsibility, should be evaluated for use. This would help arrive at a liability framework that provides flexibility on allocation while giving certainty to individuals that someone will be held accountable for any harm caused by AI. Comments on Section II.1.2 on "Data Governance"Caution must be taken with the language provided in this section to ensure that it does not contradict the provisions of GDPR. Given that data governance principles for AI will likely cover the same or similar concepts as those expressed in GDPR, it would be important for the Guidelines to approach the same or similar concepts in a manner consistent with that of the GDPR while avoiding contradictions.Comments on Section II.1.3 on "Design for all"We have no comments.Comments on Section II.1.4 on "Governance of AI Autonomy (Human oversight)"• There is a divergence of opinion here. Some argue that this should be a fundamental right with no exceptions. Others argue that implementation should be situation specific.• The Commission will need to decide on how to approach this diverging issue, and in particular, consider how governance of AI autonomy should be implemented in practice. Consider learnings from the GDPR in devising an ethical and practical implementation.• In our view, it may also be helpful to add a note on the importance of providing human oversight for detecting and protecting against cyber attacks. Further, it may be helpful to add another note stating that there must be	Comments by Baker McKenzie:Chapter III: Key Guidance for Assessing Trustworthy AIConsider the implementation of "Ethical Impact Assessments" along the lines of Privacy Impact Assessments under the GDPR. Use of the AI solution must be lawful, fair and transparent. This could be a clear process that must be carried out at the start of any AI project to quantify the major risks involved. The list provided is a good start but the requirements and questions asked must be very simple and clear so that they can be understood and implemented by non-lawyers.Suggest an assessment procedure endorsed by a recognised third party (i.e. EU entity, the experts, others). As for the software released by the French data protection authority to perform a GDPR impact assessment, would be very beneficial to have something similar for the AI impact assessment.Should the list be auditable by a regulator? Should it be sent to a regulator in particularly high risk situations? Consider the Privacy Impact Assessment model. Would this work here?Comments for (1) Healthcare Diagnosis and Treatment- Particularly sensitive data- Significant decisions - higher requirements for human oversight and final decision making- Higher requirements for auditing and explaining decisions, particularly where they may affect patients healthComments for (2) Autonomous Driving/Moving- Explicability in the event of an accident is key in being able to establish which actor should be liable- Need for a liability framework to allocate risk properly and ensure compensation is paid by the correct actor- Does this use case ("Autonomous Driving/Moving") cover AI-driven drones or unmanned aircraft systems (UAS)? It may be worth considering addition of discussion on AI-driven drones or UAS as well.Comments for (3) Insurance Premiums Comments on (4) Profiling and law enforcement- High risk of discrimination / bias within such systems- Need to mitigate systemic bias and be careful with the implementation of such systems to avoid perpetuating discrimination	-

of protecting against development and deployment of AI that could be harmful for the environment.• AI's potential to accelerate medical discoveries and to improve healthcare. Comments on sub-section on "The Principle of Explicability: `Operate transparently'" under Section I.4With regards to the discussion on "informed consent" (page 10), please see our comments provided for Section I.5.1 on "Identification without Consent" below.Relevant in the context of "transparency" appears to be the question on whether there should be a right of citizens to be informed of automated treatment of their data by government bodies and have a right to be offered the right to 'opt out' bearing in mind the fact that this type of technology will also - conceivably - be used by governmental agencies and police enforcement. Another point to consider in this context is a potential appeals mechanism (allowing a right of appeal against decisions made by AI). General comments on Section I.5 on "Critical concerns raised by AI"We agree with the Guidelines in that "[p]articular uses or applications, sectors or contexts of AI may raise specific concerns, as they run counter the rights and principles set out" in the Guidelines. The Guidelines' non-exhaustive list of critical concerns appears generally helpful in providing concrete examples of areas that are subject to higher risks of AI's unethical use and deployment (i.e., AI's breach of the "ethical purpose"). Further, it may be helpful for the Guidelines to discuss whether an ethical evaluation of each of the identified areas of "critical concerns" should also consider (1) the specific type or nature of AI technology being developed and/or deployed, (2) whether the technology is being applied in a commercial context (e.g., by companies) or in a non-commercial context (e.g. by the government), and (3) the sector and industry where the technology is going to be used. It is key that the areas of critical concern are made more specific by reference to certain use cases / factual scenarios which give rise to ethical issues. These can then serve as clear guidelines to actors in these areas. Comments on Section I.5.1 on "Identification without Consent"• The suggestion in 5.1 for a default assumption that "consent to identification has not been given" is meaningful in that it highlights the legal and ethical concerns associated with identification without consent. However, we believe it may be helpful for the Guidelines to also discuss the potential repercussions of that default assumption on the development and deployment of useful AI technologies, both from public policy and commercial viewpoints.Comments on Section I.5.2 on "Covert AI systems"• We agree with the overall principles outline in the Guidelines, but foresee some difficulties that may arise with their practical implementation. This may warrant further investigations by the European Commission and/or its High-Level Expert Group on AI. • It may be worth further exploring the scope of the requirement that AI developers ensure that humans are made aware that they are interacting with AI. Here it may be worth considering the specific context in which the AI solution is being used. We believe that there are at least the following two ways to approach this issue: (1) Being able to human oversight in not just the initial decision for deploying AI, but also for conducting systematic assessments on whether the AI's use should be continued or not. Comments on Section II.1.5 on "Non-Discrimination"• In addition to the bias that may be caused by the use of biased and/or incomplete data sets (as identified in the Guidelines), bias may also result from the introduction of inherent bias held by designers and/or developers when writing their AI algorithms (consciously or unconsciously). This is why having diverse and inclusive design teams (as discussed as one of the "non-technical methods" is important in fighting against discrimination).Comments on Section II.1.7 on "Respect for Privacy"• See Comments provided for "II.1.2. Data Governance" above.Comments on Section II.1.8 on "Robustness – Resilience to Attack"• In our opinion, it would be helpful to mention the importance of the AI system to operate robustly—reliably and consistently—not only in "normal" situations, but also (and perhaps more importantly) in unforeseen situations. This can help build people's trust in the technical robustness of their AI. • In our view, systematic assessment of the training data (e.g., checking for completeness and absence of biases) can be helpful in providing the necessary robustness. Comments on Section II.1.9 on "Safety"• Similar to our comments provided for "II.1.8. Robustness – Resilience to Attack" above, we believe it could be helpful for the Guidelines to discuss the importance of AI to operate safely in both "normal" and unforeseen situations. Mechanisms should be put in place to ensure safe operation of AI when unforeseen conditions are presented or when the AI system is under attack (e.g., to prevent harmful and unsafe use of the AI by hackers).Comments on Section II.1.10 on "Transparency"• It may be helpful for the Guidelines to additionally discuss the possible benefits of being transparent with the biases and/or training failures that resulted during the development process, if any.• Various research papers* propose a multiple prong approach of control and transparency with regard to AI solutions. The underlying thought process is that if you could/would query why a human has taken a particular decision, the same should apply to any decision being passed on an AI. Below are suggestions of how such transparency can be achieved, which we recommend the EU Commission to investigate further. • The following potential information could be required from an AI solution to provide transparency. (1) Explanations: which an AI would have to be programmed to provide to reflect its decision process (this would be reflected in a (fundamental) right of explanation). This explanation could include a human-interpretable description of the process by which an AI based decision-maker took a particular set of inputs and reached a particular conclusion. It could include a catalogue of question which have to be answered by default and automatically. For example,. "would changing a certain factor, such as ethnicity or age, have changed the AI decision" and "why did two similar-looking cases result in a different decision, or vice versa? " We note that there may be limitations to the explanation / disclosure of information about AI decision making processes due to protection of IP

'request and validate' the fact that they are interacting with an AI identity could, in some instances, be more straight forward to implement than a general awareness requirement. This would apply in particularly short engagements with AI e.g. in an automated phone system. These could be made inefficient if legal notices had to be included in each instance of such an engagement; or (2) It could equally be argued that such a right to information is based on basic fundamental rights, including human dignity, which should have no exceptions. This would support a human centric approach which would require that all use of AI, especially when it is utilised in a decision process, should be flagged. If the AI decides without any human intervention or substantial human review, citizens should be informed, even if it affects short interaction. In the telephone example given above, it can also be argued that such an announcement could be combined with the general announcement that the calls are being recorded.Comments on Section I.5.3 on "Normative & Mass Citizen Scoring without consent in deviation of Fundamental Rights"It is our view that:• Particular precautions and care should be taken with respect to AI's use for citizen scoring, because, as provided in the Guidelines, improper use of normative citizen scoring can endanger the values of freedom and autonomy of all citizens, as well as other fundamental rights, principles, and values. • It must first be acknowledged that computing any type of normative "score," especially ones that go beyond measuring physical characteristics, such as tests involving people's emotions and/or mental characteristics, is inherently prone to inaccuracies, subjective factors, and underlying biases. For example, tests for measuring one's intelligence quotient (IQ) score—which have been around for quite some time—are subject to many criticisms, including the tests' (1) inability to consider the complex nature of the human intellect, (2) underestimate of factors like motivation, emotions, and social skills, and (3) negative effects of creating biases and of polarizing the population (e.g., IQ tests drawing conclusions on intelligence based on race, gender, or other demographics). "Intelligence" means different things to different people, and its assessment involves numerous subjective factors and criteria, for which providing a score is not only difficult, but can be harmful to societies. Now, arriving at an "accurate" or "correct" normative citizen score for one's "moral personality" or "ethical integrity" would be even more challenging, if not impossible, because the assessment will likely be subject to even more subjective factors and complex considerations than assessing a person's IQ score.• Even assuming for sake of argument that an "accurate" or "correct" citizen score can be derived, extreme care and precautions must be taken in handling such information. For example, if a certain demographic (e.g., whether by ethnicity, gender, nationality, education level, religion, etc.) receives a "low" average normative citizen score for whatever reason (e.g., socioeconomic differences), that can create many negative societal effects, such as newly created and/or reinforced biases (e.g., racism, sexism, xenophobia, etc.), especially if that information is shared and/or made rights and protection of data privacy. These rights must be balanced in such disclosure exercises. We recommend raising this as a concern here. (2) Empirical Statistical Evidence: measures of an AI system's overall performance, (e.g. bias or discrimination) can be ascertained statistically. This could be considered for AI use cases where the outcomes can be completely formalised (and would warrant strict liability).(3) Theoretical Guarantees: in certain, more limited instances, AI can theoretically provide limited theoretical guarantee, i.e. for situations in which both the problem and the solution can be fully formalised. (* Literature cf Doshi-Velez, Finale, and Mason Kortz. 2017. Accountability of AI Under the Law: The Role of Explanation. Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper. Sandra Wachter, Brent Mittelstadt, and Chris Russell.) Comments on Section II.2.1 on "Technical methods"Comments on sub-section on "Ethics & Rule of law by design (X-by-design)" under Section II.2.1 • May be possible to implement AI norms into AI systems, but this will only work if the rules are very clear and easy to implement. AI HLEG will need to focus on clarity of the rules to make it possible to implement them 'by design'. • In our view, it would be helpful to add that it is important for AI's developers and designers to invest time in anticipating and establishing safeguards against unintended interactions and/or attacks. This would promote the principles, "beneficence" and "non maleficence," among others. See also our comment above on 9. Safety. Comments on sub-section on "Architectures for Trustworthy AI" under Section II.2.1We have no comments.Comments on sub-section on "Testing & Validating" under Section II.2.1 • We believe that testing and validation is an important part of AI implementation, which should be reaffirmed in the next version of the Guidelines. • In our view, this is an issue that should be considered and investigated in more detail due to its fundamental rights implications. As mentioned above in the context of flagging the use of AI, we believe that there are at least the following two ways of approaching this issue: providing strong guidance and, in due course, potentially introducing regulation that provides a best practice framework as regards to the identification of the evaluation criteria / methodology. Alternatively, leaving this to stakeholders, until the testing and validation reaches the mature status of standardisation / certification. • Overall and on balance, a combined wait and "monitor" approach may also be a viable option.Comments on sub-section on "Traceability & Auditability" under Section II.2.1 • "Traceability and Auditability" in "black box" scenarios – important to include a discussion on the possibility of explaining decisions taken by AI. An important point to consider is how widely explanations may need to be published. • Concern here regarding how far such reports should be published / just need to be kept on file? Who should have audit rights (a newly formed AI controlling authority within the EU)? New regulators in each member state to carry this out (yes)?Comments on sub-section on "Explanation (XAI research)" under Section

public. This would jeopardise people's fundamental rights in equality, non-discrimination, and solidarity, and go against the very idea of the human-centric approach.• For these reasons, extreme care and cautions must be taken when considering developing and/or deploying AI for normative and mass citizen scoring, especially if the scoring is applied beyond limited social domains.Comments on I.5.4 on "Lethal Autonomous Weapon Systems (LAWS)"• We propose adding the following: Given that LAWS have global implications capable of affecting securities of different states, promoting cooperation and open discussions on a global scale is instrumental in finding the right safeguards against inhumane deployment of LAWS.

II.2.1We have no comments.Comments on Section II.2.2 on "Non-Technical methods"Comments on sub-section on "Regulation" under Section II.2.2 We appreciate that the "second deliverable" may provide a discussion on any need for regulation(s) to be revised, adapted or introduced (as stated in the Guidelines on page 21). The following are proposed additions to the Guidelines if they have not yet been considered by the Expert Group:• It would be helpful to conduct systematic assessments on (1) AI's impact on and interplay with various laws, (2) whether the use of AI has raised new legal questions that must be addressed, and (3) any need to revise or adapt existing regulations or to introduce new regulations.• Regulatory decisions must weigh the potential benefits from the deployment of a certain AI system against the risks that may be caused by the deployment of that AI system. Comments on sub-section on "Standardization" under Section II.2.2We have no comments.Comments on sub-section on "Accountability Governance" under Section II.2.2• In addition to appointing a person, panel, or board that provide oversight on ethics issues, as suggested in the Guidelines, we believe that accountability governance may be further strengthened by designating person(s) to assume responsibility for actions taken by AI that result in harm. This would help realise Trustworthy AI. See above comment on this. It seems difficult to identify within an organization who should be accountable for the development or deployment of AI, where these are imputable to the organization as such. The normal ruled on accountability should be applied (civil, criminal, administrative liability). Creating a parallel accountability regulation for AI seems counterproductive for the fostering of AI.• As referred to above, the issue of allocating liability is not easy to solve due to the complex nature of relationships between AI developers and AI users. May be best to consider joint responsibility, details to be defined on a case by case basis. This would provide some flexibility on allocation while giving certainty to individuals that someone will be accountable for any harm.Comments on sub-section on "Codes of Conduct" under Section II.2.2 These are key facilitators for building trustworthy AI, since codes are drafted by stakeholders directly and therefore they represent effective instruments for the market. Would propose that codes are approved by a recognised third party (which may be EU board, the Expert and/or others). Comments on sub-section on "Education and awareness to foster an ethical mind-set" under Section II.2.2• We agree that education plays an important role and that adequate education must be provided to people making the products, the users, and other impacted groups. In our opinion, the significance behind providing sufficient education to the people making the products (e.g., designers and developers) has been sufficiently explained in the Guidelines. On the other hand, however, there is no discussion of the risks that arise from not providing sufficient education and awareness to the users (e.g., companies or individuals) of AI. • For the reason provided above, we propose adding a brief discussion on the potential harm that may result from improper education and/or insufficient

				awareness among AI users. For instance, if the users do not fully understand or appreciate the limitations of their AI systems, then that may result in their overreliance on the AI systems; for example, if a user of a predictive AI system blindly or overly trusts a prediction made by the AI system, particularly over a more well-supported prediction made by a human, then that may have unfair and/or other negative consequences. This would go against several ethical principles set forth in the Guidelines, such as “beneficence,” “non maleficence,” “autonomy,” and “justice.” • It is therefore important for AI users to be educated and made aware of the limitations and weaknesses present in the AI systems they are using. In other words, AI’s “decision-making” should not substitute humans’ decision-making. Instead, AI’s capabilities should merely be used to help or complement humans’ decision-making processes. Comments on sub-section on “Stakeholder and social dialogue” under Section II.2.2• It may be helpful to add the following to the exemplary list of “experts and stakeholders”: government representatives; industry representatives; and various subject-matter experts.• Further, it may be helpful for the Guidelines to expand on its description of ways for the general public to engage in discussions that can effect meaningful change.• It may be worth adding a note mentioning that sharing of best practices by the various stakeholders is one way of promoting effective open discussions on AI.Comments on sub-section on “Diversity and inclusive design teams” under Section II.2.2• It may be helpful to mention that, as discussed in other parts of the Guidelines, bias and discrimination may result from, among other things, inherent biases held by AI’s developers and designers, as well as from the use of incomplete and biased data sets as training data. Having diverse and inclusive design teams can help in (1) designing AI systems that are less biased (and more objective), (2) identifying and using less biased and more complete data sets, and (3) finding ways to offset the negative effects from biased training data. • Further, as already outlined above, having diverse and inclusive design teams may also help AI’s developers and designers in identifying—and protecting against—possible situations that may result in discrimination of people. Comments on “KEY GUIDANCE FOR REALISING TRUSTWORTHY AI” under Section II (p. 23)Since information and traceability requirements would be the same as under the GDPR, it may be more straightforward to encourage stakeholders to extend to all AI projects the same measures already in place for data protection.	
Steven	DEWAELE	Huawei Technologies		Page 17, Line 23, on Reliability & Reproducibility With some online adaptive learning systems, the system model parameters (can) change over time. In that case, the same input will a fortiori produce different outcomes at different times. In theory, you can always bring back the system to the state it was in at the time the questioned/contentious result was produced.	Page 27, Line 16, on Transparency and Traceability When it comes to traceability, the guidelines suggest to provide documentation on the method of building the algorithmic system. Huawei is happy to support this guideline, but believes it is advisable to specify that the guideline is obviously not meant to compromise any company's intellectual property rights.

					But in practice, this may prove to be more difficult. It would require such a system to store all the data it has ever used to train itself, or to continuously store copies of itself.			
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	
Daan	Kayser	PAX - on behalf of the Campaign to Stop Killer Robots		The Campaign to Stop Killer Robots appreciates the initiative taken by High-Level Expert Group on Artificial Intelligence in drafting the AI Ethics Guidelines for Trustworthy AI. Our campaign was founded in 2012 and currently consists of 89 civil society organisations from over 55 countries. We appreciate the fact that the guidelines address the issue of lethal autonomous weapon systems (LAWS), as this is a fundamental issue in the debate on AI. Definition and risksThe Draft Guidelines define LAWS as weapons without meaningful human control over the critical functions of selecting and attacking individual targets. This approach to definition is welcomed, however it would be better if the text recognised that such systems are a developing concern. Thus it should say LAWS "would operate", rather than LAWS "can operate".The guidelines usefully point out some of the fundamental ethical concerns, the risk of the emergence of an arms race, as well as the issues of system malfunction and the risks of military contexts with no human control. However, the guidelines could further explain what other challenges there are related to LAWS. Although the guidelines state that human beings must remain responsible and accountable for any casualties, such systems raise a fundamental ethical concern that life and death decisions must not be delegated to a machine. Beyond this, we would see an erosion of international legal frameworks if the frequency and nature of human legal judgement is allowed to be diluted. Meaningful human controlIn addition, we suggest that the section on LAWS would refer to the concept of 'meaningful human control' as the central element in the debate, which should be the fundamental principle that guides the development of a legally binding instrument prohibiting LAWS. The positive obligation of meaningful human control steers us away from complex debates about the definition of autonomy and related technology, and would remain relevant irrespective of unanticipated technological developments.Collateral damageThe guidelines mention that "LAWS can reduce collateral damage, e.g. saving selectively children." This is an entirely speculative statement, one that most experts in artificial intelligence would refute and that is backed up by no evidence. Whilst new technology may have the potential to reduce risks to civilians, this text is specifically concerned with LAWS as systems operating "without meaningful human control." There is no basis for asserting that systems operating without meaningful human control will reduce harm. Furthermore, "saving selectively children" appears to be an abandonment of the protections afforded in international law to all other people who should not be attacked.			

This claim should be deleted.The European Parliament’s ResolutionWe are pleased to see the AI HLEG “stands with, and looks to support” the European Parliament’s Resolution 2018/2752(RSP).The guidelines refer to the European Parliament’s Resolution urging the development of a common legally binding position addressing ethical and legal questions of human control, oversight, etc., but the guidelines do not refer to the crucial paragraphs 2, 3 and 4 of the same Resolution. Notably, paragraph 3 states that the Resolution “[u]rges the VP/HR, the Member States and the Council to work towards the start of international negotiations on a legally binding instrument prohibiting lethal autonomous weapon systems”. Paragraph 4 stresses that “in this light, the fundamental importance of preventing the development and production of any lethal autonomous weapon system lacking human control in critical functions such as target selection and engagement”. Thus, the Resolution is not only about ensuring a common position, but also about taking pro-active steps towards preventing the development of such weapons and working towards banning them. Also the resolution refers to a common position “that ensures meaningful human control over the critical functions of weapon systems, including during deployment.” As this Resolution has been adopted by the European Parliament by a large majority, we believe it would be good to see the contents appropriately reflected in the updated Ethics Guidelines. Societal concernBesides the European parliament resolution, there have been urgent calls from numerous parts of society warning for these weapons and advocating for a prohibition of these weapons. To demonstrate wide societal concern it could be good to mention these. Recently the UN Secretary-General called for a ban, calling the weapons “politically unacceptable and morally repugnant”. The ICRC recently stated “limits are necessary for addressing legal, ethical and humanitarian concerns” autonomous weaponsIn 2015 over 3900 Artificial Intelligence experts, and in 2017 116 CEO’s from robotics companies warned against these weapons and called on the United Nations to take action. In 2018 240 tech companies and over 3000 individuals pledged to never develop, produce or use of lethal autonomous weapon systems.In sum, we welcome the recommendations made by the Draft Ethics Guidelines for Trustworthy AI. The guidelines appear to address multiple concerns raised by the advances and proliferation in AI techniques. Nevertheless, we recommend the inclusion of the additional points we raise above, particularly in reference to Lethal Autonomous Weapons Systems.

Silvia	Elia	Consorzio Netcomm - Italian Consortium of Digital Commerce	Netcomm agrees with the observations raised from the European Commission and the High-Level Expert Group on Artificial Intelligence (AI HLEG), welcoming the structure of the guidelines: the three areas of analysis will allow to examine in details the phenomenon, identifying the most important and crucial aspects. Indeed, it is fundamental to approach the develop of A.I. considering all the aspects involved with the awareness that the fluidity of the current social, economic and technological context requires to reconsider it periodically, at the light of the technological developments as it does not allow, today, to consider all the possible scenarios.The complexity of the scenario requires a systematic approach as it involves aspects that are profoundly different but inevitably connected between them, such as technology, ethics, regulation, economy, regulation, government, etc. From this assumption derives also the deep difficulty of interfacing with the relevant Stakeholders for each sector to satisfy all the needs.	Netcomm fully agree with all the remarks developed by the A.I. HLEG in terms of "Fundamental Rights of Human Beings" and "Ethical Principles in the Context of AI". To determinate the framework in terms of to set of principles, values and purposes is crucial, as well as the possible risks that could arise in a long-term period, represented the key for minimise the dangers benefiting of the positive effects.Referring to Chapter I), Section 5, "Critical concerns raised by A.I.", Netcomm asserts that the most important points of attentions have been touched by the AI HLEG, furthermore, Netcomm would submit some further considerations.Referring to Chapter i) Section 5.1.). Through connected devices (such as smartphone) and services, companies are already able to collect data in every industry. In particular, the most advanced digital marketing systems process huge quantity of information every minute, often without making aware the interested parties.Another critical area – where a huge mass of data is generated and processed, is Transport sector: for example, the rapidly increase of sensors and cameras allow to collect massive data (i.e., booking services or security systems into the critical infrastructure such as airports or train stations). Related to these issues, in addition to the findings raised by the Group, Netcomm highlight other aspects that must be considered as the increasingly difficult to determine the criteria of attributing ownership and legitimate the use of the data and related responsibilities. Referring to Section (5.2.). Netcomm notes that Citizen Scoring activities could become more pervasive with evident risks on the fundamental rights, especially for some categories such as vulnerable people (such as minorities or disabilities) or in situations where there are clear asymmetries of power. Relating to this point, it should be noted that even Article 22) of GDPR seems to not cover all the possibilities and implications that could be generated. The Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 give important prospective to face these issues but, as anticipated, are not fully sufficient to face all the possible needs. We underline that profiling processes and the related services are becoming more and more widespread and required both by companies and Public Administration. The profiling activities and the knowledge of the customer have become the fundamental starting point for determining the success of the business strategy for companies. Netcomm recommends investigating deeper these aspects.Regarding to Section (5.5.) Netcomm believes that "Potential longer-term concern" lays down under the point of Responsibilities. The A.I. Systems mentioned by the AI HLEG – (i) AI systems that may have a subjective experience; ii) Artificial Moral Agents; and iii) Unsupervised Recursively Self-Improving Artificial General Intelligence (AGI)) – could rise some of the most critical aspects in terms of civil and criminal liability of algorithms and advanced machines as well as jurisdiction issues.Until now the direct responsibility of the machine has been peacefully rejected – considering that in the most part of the cases it was attributed to the producer and, in limited cases, to the developer – nonetheless, for the future development it cannot be	No comments to add.	3) Analysis and consideration on Chapter 3Netcomm would submit its point of view about the Assessment List.2. Data governance.Netcomm agrees with the remarks but would like to underline the need to better define the processes, the criteria to state (human) responsibility on all levels of the treatment (i.e. both the develpoment than use of algorithms) and the criteria to value the legitimacy for the use of the data; further questions that could rise, such as: "How could we determine the Jurisdiction in case of damage caused by incorrect data handling? Should the GDPR criteria are sufficient?"4. Governing AI autonomy:Netcomm totally agrees with all the assessments listed, believing that this is one of the most important and crucial point; Netcomm adds that it could be helpful, in certain circumstances (especially where autonomous robots are allow to interact in delicate areas) to set an internal supervisory mechanism whose members are external, nominated periodically, composed by technicals and experts from other areas involved (such as lawyer etc.) in order to value not only the first level of the process (such as the algorithms) but also the results, i.e. the output that could be generated from the machine (which it could not therefore be previously determined). 6. Respect for Privacy.Netcomm supports and agrees with all the points; it adds that data protection is part of human rights and for this reason its protection should be applicated independently from GDPR, being an instrument to implement this protection (it means that evaluating simply compliance with it seems reductive respect the values discussed). The developers should place in the centre of their reflections the protection of human dignity. In any case, GDPR recognises fundamental rights for users, it is therefore necessary to guarantee the respect of these rights even when the treatment is carried out by machines.7. Respect for (& Enhancement of) Human Autonomy.Netcomm totally agrees and adds some considerations. It could be useful, in certain circumstances, to set out verification procedures that could reply to such questions: - "Are there mechanism or systems that allows to the user to submit easily complain to the owner of the process the machine/algorithm decisions in case of (alleged) prejudice?"- "Is the machine/algorithm decision binding for the users?" - "What are the instruments (included organisms) that are authorised to analyse the issue?"10. Transparency.Netcomm totally agrees with these remarks and would underline the need to guarantee transparency of the processes, especially in certain sectors or areas where the human being could be prejudiced. Processes should be scalable and verifiable on all level by independent and third-party assessment systems. Therefore, transparency should include:- Processes;- Liability for each level of the process;- Assessment;- Traceability of the data and the output.- Mechanism to withdraw date, information and destruction of the most sensitive information.Consideration about e-Commerce, Digital Marketing and Profiling.Finally, following the recommendation of the HLEG to share comments about one of the areas mentioned in the documents, we would close this contribute with some last considerations	Netcomm, the Italian Consortium of Digital Commerce, is the reference point for e-commerce and digital retailing at national and international level. The Consortium aims to promote the spreading of e-commerce and the digital evolution of companies, thus generating value across the entire value chain and consumers.Netcomm believes that Artificial Intelligence Systems would have a strong impact on the global scenario, for businesses and citizens, becoming a key driver of economic development. The strategy adopted by the European Union moves forward into the right direction, promoting technological developments in order to compete with other world-powers and, in the same time, taking in the right consideration all the ethical and social aspects involved.Netcomm, as expert of e-Commerce and digital Sector, would give its contribute submitting the above consideration on the Ethics Guidelines remaining available to support the Group of Experts in the development of their reflections.

excluded that the evolution of this sector will lead to previously unexplored issues of law.In particular, the following issues can be raised: i) How qualify automatic systems under the legal aspect; ii) How to attribute responsibility if an AI (or machine learning) system causes damage not planned by the user and not foreseen by the developer? iii) Is the machine imputable under the criminal law? iv) Is there any responsibility for the developer even in the event of an error, defect or malfunction of an intelligent robot? More questions could rise. Another aspect is relating to the Jurisdiction, the first question that must be asked is: "How to determine the Court jurisdiction in the case of damages determined by machine learning?"; "Could the current criteria be used?".Furthermore, additional legal and economic aspects could be mentioned; although they do not generate serious and direct consequences for the human being, they could nevertheless determine legal issues of considerable economic value.Netcomm refers to the issues related to the Intellectual Property. Until now, only the aspects relating the development of computers and algorithms have been considered but, in the future, new and unexplored issues could arise, especially under the legal point of view. It means that it will be necessary to reflect on how to qualify and resolve possible controversies that may arise in relation to the intellectual works produced directly by the machines.

about the impact of A.I. on e-Commerce and Digital Marketing.Netcomm believes that systems based on A.I. would have a positive impact on the scenario, for businesses and users but at the same time they are opening important questions.Indeed, A.I. could be decisive to cut down some serious issues that are afflicting e-Commerce, as fraud detection and prevention for online transactions, to contrast counterfeiting of goods, especially in the pharmaceutical or childcare sector, as well as identity theft, etc. Technologies like blockchain will allow users to have more control of their data thanks to greater transparency in management, the power to control of flows, protection from threats, malware, from the risk that data end up in the deep web.A.I. could improve services and user experience in terms of search, features and personalization, contributing to recommendation and purchase predictions tailored on the users, or to develop predictive customer service. However, awareness of the unpredictable is also required.Digital Marketing is one of the fields in which automatization and machine learning would open new scenarios with solutions tailored on users but also unpredictable implications.Currently, the new frontiers of digital marketing use new technologies such as systems of virtual and augmented reality which are opening up new and more advanced profiling activities.We are already witnessing the first case histories that use systems based on blockchain technology that allows to record in a more effective and precise way the customer journey along the entire journey and his/her experience experienced by the consumer.The system becomes the collector of data across all touch points: this means that all the user actions and interaction are acquired and stored: from the opening of the email and newsletter, to the registration on website and following accesses or the app's download; from the purchase of the product online or offline to payment, etc.All these actions are recorded on the ledger, validated with a certain date and made unchangeable, thus attributing certainty to the identification of the person, his actions and his preferences. Not only (already now) these systems are able to react to events in the real world and are able to grasp - thanks to the acquired information assets - not only rational behaviours and interactions, but also irrational and unconscious behaviours. We can not therefore exclude that soon they will also be able to guide them (humans).The most critical aspects lay down on this point. The questions might, moreover, be: "To what degree is it possible to regulate these phenomenons?", "How far does the algorithm (and the learning machines) could be push forward?", "How far does predictions and direct and indirect conditioning of human actions and thoughts be pushed forward?".

Katherine	O'Keefe	Castlebridge	The European motto “united in diversity” serves as an important reminder that for ethical use, AI must consider whose voices and whose power are being represented in development. Which humans are being centred in human-centric AI? Whose definition of the common good? Trust is essential, and as observed, trustworthiness is an outcome of quality data processing for good. Trustworthiness is a quality characteristic of data, but as a characteristic it always raises questions of perception (perception of confidence in quality), and of expectations. (Trusted to do what? Trusted by whom?) It serves as a reminder that we must never forget that while "ethical purpose" and "robust technology” are important, we must always consider the context of our intent, purpose and development, and must consider possible effects in the social context of application.As such, we would re-emphasize the importance of paying particular attention to asymmetries of power or information and situations involving vulnerable groups as stated in the guidance on ethical purpose, and suggest that the very act of developing AI is likely to involve a power imbalance. The relationality is extremely important to consider.We note that the document specifically states that guidance does not replace legislation or regulation and look forward to dialogue on what will be required to enforce development along ethical guidelines.	We would query the formulation of Section 2. “From Fundamental Rights to Principles and Values”. While we recognize that the section does complicate its own formulation of an apparent linear relationship of Rights to Principles to Values, we would suggest that the relationship as discussed may be the other way around; that fundamental rights are a concrete expression of concepts to realize formulated principles such as autonomy and equality, which express our values, or what we hold to be good or of importance. In regards the example of “informed consent”, “Value” is an imprecise term for the construct needed to express principles and uphold rights. “Informed consent” is not a value, but a mechanism to control operations in line with principles that express our values.We acknowledge the centrality of a Rights based ethical framework to the guidance presented but suggest that this might be broadened. The Rights based focus underpinning the guidance presents a clear but limited approach. While using the rights-based approach is consistent as a basis for legislation, it would be useful to include in non-legislative guidance approaches that situate the intended development and use of AI in frameworks that interrogate relationality and outcomes as well as presenting abstract principles. Issues of distributive justice are often less clear in the rights-based approach, than, for instance, an Ethics of Care. Additionally, while the rights-based principles described have presented a somewhat international convergence point in law, the convergence often covers differences in cultural interpretation and operational limits. We need to avoid a restrictively classical Western approach to be able to, for instance, find a common ground with Asian ethical frameworks and uphold indigenous data sovereignty. Ethical Purpose as a concept suggests intent, which by its nature reminds us that purpose and intent are only part of the concern. Development of AI must be alert not only to ensure ethical purpose but to ensure design that protects against abusability. Consideration of abusability reflects the risk-based approach inherent in Privacy by Design under GDPR, and Ethics by Design by extension. This requires a consequentialist approach in considering ethical impacts as well as operational alignment with principles. While this may be difficult in considering long-term potential effects, difficulty should not prohibit consideration. In regards to Justice/Fairness as a principle: Consideration of stakeholder needs or the needs and requirements of different segments of the population must be considered in questions of fairness and equality. Equality of access or equal treatment in a system where the good does not equally meet the needs of different populations may appear to be fair but cover a deeper injustice. In developing AI, this requires deeper questioning regarding the context of design projects in society. This is not only a question of equality of access, but equality of benefit.In this, we would specifically emphasize the importance of considering asymmetries of power or information as mentioned in the draft guidance and expand that to asymmetries of voice in input and design.	We would note that the described systems development process (in particular the figure illustration) does not appear to include returning to first principles as a quality assessment process: “Does the outcome uphold or violate the core principles?” Do the methods for implementation succeed in ensuring that the developed process does so? Quality Feedback assurance must go back to validate against principles and values. For reference, see the Impact Assessment Model on p. 263 of O’Keefe and O Brien, Ethical Data and Information Management. Kogan Page 2018, influenced by Data Quality models by Danette McGilvray and Dennedy and Finneran’s Privacy Engineer’s Manifesto.Realizing principles in operation requires measurable quality characteristics. We have suggested the following quality characteristics of ethical information management outcomes in addition to impacts on human rights and freedoms. Some of these “quality characteristics” directly reflect the principles elucidated earlier in the guidance:Utility: A measure of to what extent the information and/or process outcomes will do good in society or will promote happiness. This quality follows directly from Irish philosopher Francis Hutcheson’s definition of the principle of utility: ‘that action is best, which procures the greatest happiness for the greatest numbers’ (Hutcheson, 1726).This characteristic is very broadly defined but may at the same time be useful for its broadness of definition. This characteristic is one of the dimensions of the Castlebridge utility/invasiveness model used as a brainstorming tool for considering impacts.Some metrics you may look for are stakeholder satisfaction, the degree to which your process or outcome solves a problem, etc. At a more operational and functional level, you can also look at this characteristic as a measure of ‘usefulness’ to individuals or to society.Beneficence/non-maleficence: A measure of the extent to which the process or processing promotes well-being, or the extent to which the processing supports physical well-being and the good of society in a way that doesn’t cause harm. (This clearly reflects principles listed earlier in the guidance)Justice/fairness: Justice is a clear expected outcome for ethical information management. In this context, justice and fairness can be defined as a measure of the extent to which your processing results in equal treatment of people or even increased equality. Information outcomes and process outcomes that rank strongly on the dimension of justice/fairness will result in the equal and fair treatment of people, results or distribution of resources. Those that do not will result in some curtailment of equality, some bias against individuals, and some unfairness in the distribution of resources. This quality characteristic is a key metric that has been identified in questions of algorithmic accountability. Disproportionate impacts on vulnerable or marginalized populations are essential to consider.Verity/non-deceptiveness: Verity as a data characteristic relates to the integrity, truthfulness, honesty or accuracy of your representation, construction, or the results of information management. It is best defined as a measure of how closely your processing activities and use of data match what you had declared your processing to be. Verity/non-deceptiveness is an external	Detailed guidance will need to address situational modifiers. From a practical perspective, guidance will also be needed to support decision making not only regarding sector specific contextual settings, but in legislative context and considering requirements to support whistle blowers.	The work of various initiatives and scholarship speaking from the marginalized to power such as Data for Black Lives and Indigenous Data Sovereignty initiatives such as the Maori data sovereignty network would be valuable to inform guidance for specific use cases, and to consider the power relationships, implicit and expressed, in publication of this guidance.

					corollary to traditional information quality metrics such as McGilvray's internal 'quality of information specification' or 'perception, relevance and trust'.Autonomy: the measure of the extent to which the outcome of your process respects or infringes on people's self-determination or ability to choose an action for themselves. This measure is influenced by the extent to which people are able to make their wishes known in relation to the processing, whether the design of a system or process is transparent in allowing choice or whether it suppresses true, informed choice. An information imbalance where an individual 'agrees' to something they are not aware of agreeing to does not represent a true choice, which constrains their autonomy and could also be considered a defect in the context of the verity/non-deceptiveness dimension. Likewise, obfuscation through information overload – that makes it harder for people to understand their choices – also can constrain autonomy.In the broader context, processing that removes the potential for choice from an individual by, for example, not making information about a product, service or other benefit available to them based on the processing of data about them, or which results in constrained choices for that person in the exercise of other rights or freedoms, would also be processing that would impact on autonomy.Privacy/invasiveness: Privacy/invasiveness is a measure of the level of intrusion in to the personal life, relationships, correspondence or communications of the individual or a group of individuals as a result of the processing activity or the information outcome or process outcome that is delivered. It is not a measure of compliance with privacy laws, although this may be a factor you might consider in an analysis. One aspect of privacy/invasiveness is the level of autonomy or choice that an Individual can exercise over the processing of data about them.Necessity: is a measure of the extent to which the proposed processing is addressing an issue that, if left unaddressed, may result in harm to or have some other detrimental effect on society or a section of society. This is based on the analysis of the collective body of the EU's Data Privacy Regulators.Proportionality: Proportionality is best defined in this context as a measure of the degree to which the interference in privacy, and the potential infringement or curtailment of other rights, caused by the measure is counterbalanced by the benefit to society or a section of society arising from the objective being pursued.A key ethical test is to determine if the same objective could be achieved with a more limited impact on individuals and their autonomy or other rights. (pp. 214-218. O'Keefe and O Brien, 2018)	
Julian	Stubbe	Institute for Innovation and Technology (iit), Berlin	We appreciate the rationale of the document as a "living document". It is not the right time to establish binding regulations for AI, which could even hinder the development of AI for societal benefits. For example, using health data is often considered ethically problematic in general terms, but specific usage of health data is considered good. In this regard, we miss discussing the question: what is the best scope for ethical guidelines? The discussion should address	We think, it is right to embed the discussion of "AI Ethics" within more general shared European rights and values. It is right to explain that general ethical principles can and should be transferred into the AI context. Chapter 5: The topics addressed here seem to be very selective. What about concerns like: - manipulative information - stabilizing unequal power relations (i. e.	We think, the list of requirements for Trustworthy AI is sufficient. We would like to propose an additional non-technical method: - Innovation Funding: The engagement of institutions is not limited to regulation or standardization. They can also encourage research and development in AI technologies that fit the idea of Trustworthy AI. The aim should be to make Trustworthy AI a driver of innovation.	

			how to identify domains that share similar ethical implications of AI and how these domains should be differentiated: by technology (i. e. pattern recognition, deep learning etc.) or area of application (i. e. health, mobility etc.) or by something else (i. e. affected societal groups)? Such differentiations would allow the discussion to become more specific and hands on.	when AI becomes a key enabling technology, big companies have a huge advantage and may increase their power) - unequal access to AI (i. e. citizens who already possess technology benefit even more, whereas people without access to technology might increasingly lack behind) The list could go on. The crucial problem is that the scope of the mentioned concerns is unclear. Are these supposed to be general concerns that come up whenever AI is implemented? If so, concerns like "LAWS" or "citizen scoring" don't match the list, because they are application based, whereas "covert AI systems" or "identification without consent" are more generally connected to how AI works.		
Swee Leng	Harris	The Legal Education Foundation	The approach to law taken by the Guidelines risks under-valuing the importance of law and access to legal remedy to enforce the law as a vital aspect of the rule of law. Without access to an effective legal remedy, principles can be ineffective and the rule of law absent. The potential shortcomings set out in other sections could be addressed by framing the Guidelines with a legal frame, rather than an ethical one.Fidelity to rule of law principles and implementation of administrative law standards should guide government's use of data processing. The Rt Hon Dominic Grieve QC MP, recently observed in a talk at The Law Society of England and Wales: 'How are we going to operate these systems in a way where they can be challenged if the decisions they make are unfair?' Grieve asked. He noted that the Windrush scandal illustrated the risk of bureaucratic mistakes. 'If on top we are now going to factor in algorithms we are going to have to ask ourselves questions about what information are citizens going to be given, on data accuracy,' he said. While automation has the potential to transform government for the better 'it is also possible to see how it has the capacity to act very badly indeed'. When government decisions are made about individuals' rights using data processing, it can be hard for an individual to know whether their data were accurate or processed correctly. Examples set out below of problems with data processing by the UK government illustrate the importance of rules for data processing by government that promote good administration. The rule of law should still be upheld in the digital age by digital government, i.e. government that uses data processing for government processes and decisions. This includes upholding rule of law principles such as: law must be accessible and so far as possible, intelligible, clear and predictable; and ministers and public officers at all levels must exercise the powers conferred on them in good faith, fairly, for the purpose for which the powers were conferred, without exceeding the limits of such powers and not unreasonably. In digital government, the decision-maker and what constitutes a government decision can be obscured by online interactions and automated data processing. There is a risk of well-established administrative law principles not being upheld—e.g. that a government decision-maker should not be biased, and should consider all relevant considerations—when government uses algorithms and application programming interfaces (APIs)	While Chapter I recognises the importance of law and rights, many of the principles identified as ethical principles would be better understood as principles that are or should be established in law and legally enforceable. The same is true of the "Requirements of Trustworthy AI" identified in Chapter II. The following discussion sets out this point with respect to the particular example of welfare in the UK.The UN Special Rapporteur on Extreme Poverty and Human Rights recently visited the UK and the move to increasingly digital approaches to welfare was a focus of his work, from which he concluded that government use of automated systems needs more transparency and the application of the rule of law. The Special Rapporteur made a long statement at the end of his visit setting out his initial findings and conclusions on the impact of 'a digital welfare state' on human rights, particularly in relation to vulnerable individuals. The UN Special Rapporteur found that a 'digital welfare state' is emerging in the context of transformation across government with government services becoming 'digital by default'. Universal Credit is the first major government service to be digital by default. He observed:"Automated BenefitsWhile Universal Credit is a very visible example of digital transformation, an even more significant digital change is happening within the walls of central and local authorities. The merging of six legacy benefits into one new Universal Credit system aimed at reaching millions of UK citizens is in fact a major automation project. The collection of data via the online application process and interactions with the online journal provide a clear stepping stone for further automation within DWP.One example is the Real Time Information (RTI) system, which takes HMRC data on earnings submitted by employers and shares it with DWP, which in turn uses this data to automatically calculate monthly benefits. As DWP explained to the Special Rapporteur, Universal Credit is only possible because of the automated calculation of benefits via RTI.But with automation comes error at scale. Various experts and civil society organizations pointed to problems with the data feed, including through wrong or late information transmitted by employers to HMRC. According to DWP, a team of 50 civil servants work full-time on dealing with the 2% of the millions of monthly transactions that are incorrect. Because the default position of DWP is to give the automated system the benefit of the doubt,	This section looks at the settled status application process for EU citizens in the UK as an example of why the rule of law is necessary to achieve trustworthy AI, and why without the rule of law principles such as transparency and accountability cannot be realised.The settled status application process for EU citizens and their families in the UK could be an important precedent for government decision-making using algorithms and data matching. The Home Office plans to have an access arrangement with DWP and HMRC as part of the application process. In order to assess whether an applicant has been resident in the UK, part of the application process will be 'automated checks' of HMRC and DWP data. The Home Office's Statement of Intent on the application process for settled status states that:"Where possible, the application process will help the applicant to establish their continuous residence here and whether it amounts to the five years generally required for settled status, on an automated basis using data held by HM Revenue & Customs and in due course also the Department for Work and Pensions….where the applicant is an EU citizen and the automated checks of HMRC and DWP data indicate that they have been continuously resident in the UK for a period of five years, they will be granted settled status (indefinite leave to remain), subject to criminality and security checks. We expect that, for the majority of EU citizens who are or have been working, we will be able to help them confirm their residence in this way.Where the automated checks of HMRC and DWP data do not indicate that the EU citizen has been continuously resident in the UK, or indicate that they have been continuously resident here for a period of less than five years, the applicant will then be able to upload documentary evidence of their continuous residence." The Home Office has not yet provided much additional information on these automated checks. For example, the Home Office has not yet publicly explained how data will be matched between departments (Home Office, HMRC, and DWP). Data matching issues such as name changes due to marriage, or misspelling due to past administrative error could affect the integrity of the system.The answers generated by the departments will be government decisions, i.e. an automated decision, although the ultimate 'decision maker' for whether an applicant has secured settled status will be a Home Office official. The DWP and HMRC decisions can be	Some work to apply administrative law principles to automated decision-making has begun, and has identified a number of aspects of data protection law that could be used in the public law context. The High-Level Expert Group may be interested Cambridge University's Dr Jennifer Cobbe's paper on 'Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making'. Dr Cobbe's analysis highlights the potential interrelationship between data protection law and administrative law, for example she concludes that there are:"several key questions to be asked when determining whether public bodies have made an error of law in using automated decision-making. Where the decision concerns a natural person, the first question is whether or not the decision-making in question is caught by the prohibition contained in Article 22 GDPR – i.e. does it use solely automated decision-making and does it produce legal or similarly significant effects on the data subject. If the decision-making is caught by the prohibition then it should next be considered whether any of the applicable exemptions have been met. In doing so, the court may need to have regard for whether the claimed legal bases for the processing or decision-making have been met (including questions of valid consent, necessity, and proportionality, where appropriate) – where they have not, meaning that there is no valid legal basis, the public body will not have met any of the exemptions to the Article 22 prohibition. The court should also have regard for whether there exist suitable safeguards to protect the rights, freedoms, and legitimate interests of the data subject. Where the court determines that the Article 22 exemption doesn't apply, it should proceed to consider whether the processing involved in making the decision has a valid basis in law (again, including questions of valid consent, necessity, and proportionality, where appropriate). If at any point the public body fails these tests, then they lack a legal basis for their automated decision-making and the court should make a finding that they have made an error of law and thus acted ultra vires." On this analysis, the data protection law requirements of consent, necessity and proportionality will determine the legality of automated decision-making by government. Such requirements of necessity and proportionality echo existing public law in the UK in the context of the Human Rights Act, and are not novel tests of the legality of government activities.There is some force to

for government processes, and when 'the computer' or 'the system' are assumed never to make a mistake. Rather than focusing on ethics, the High Level Expert Group should focus on law as experience in the UK demonstrates. The UK government's approach to governance of data processing has focussed on ethics rather than the rule of law. The frame of ethics has been supplemented by data protection law, but this frame has failed to focus minds on the need for government to meet its administrative law obligations and standards of judicial review when using data processing for government functions. For example, the Department for Digital, Culture, Media & Sport (DCMS) Guidance: Data Ethics Framework includes as principle 2: 'Be aware of relevant legislation and codes of practice', highlighting a number of areas of law including on equalities and anti-discrimination, but fails to expressly mention administrative law or judicial review principles.

claimants often have to wait for weeks to get paid the proper amount, even when they have written proof that the system was wrong. An old-fashioned pay slip is deemed irrelevant when the information on the computer is different." The UN Special Rapporteur concluded that lack of transparency was a major issue with the government's development of new technologies, so that the existence of automated systems in government almost unknown. Civil society relies on Freedom of Information (FOI) requests to find out information on the government's automated systems, but such requests are not necessarily successful and are refused for reasons such as the commercial interests of contractors or intellectual property protections. He made the following arguments for transparency and the application of the rule of law to government use of automated systems:"But it is clear that more public knowledge about the development and operation of automated systems is necessary. The segmentation of claimants into low, medium and high risk in the benefit system is already happening in contexts such as 'Risk-based verification.' Those flagged as 'higher risk' are the subject of more intense scrutiny and investigation, often without even being aware of this fact. The presumption of innocence is turned on its head when everyone applying for a benefit is screened for potential wrongdoing in a system of total surveillance. And in the absence of transparency about the existence and workings of automated systems, the rights to contest an adverse decision, and to seek a meaningful remedy, are illusory.There is nothing inherent in Artificial Intelligence and other technologies that enable automation that threatens human rights and the rule of law. The reality is that governments simply seek to operationalize their political preferences through technology; the outcomes may be good or bad. But without more transparency about the development and use of automated systems, it is impossible to make such an assessment. And by excluding citizens from decision-making in this area we may set the stage for a future based on an artificial democracy.Transparency about the existence, purpose, and use of new technologies in government and participation of the public in these debates will go a long way toward demystifying technology and clarifying distributive impacts. New technologies certainly have great potential to do good. But more knowledge may also lead to more realism about the limits of technology. A machine learning system may be able to beat a human at chess, but it may be less adept at solving complicated social ills such as poverty.The new institutions currently being set up by the UK government in the area of big data and AI focus heavily on ethics. While their establishment is certainly a positive development, we should not lose sight of the limits of an ethics frame. Ethical concepts such as fairness are without agreed upon definitions, unlike human rights which are law. Government use of automation, with its potential to severely restrict the rights of individuals, needs to be bound by the rule of law and not just an ethical code." This call for transparency and accountability by the Special Rapporteur is consistent with Principle 6 of the DCMS Data Ethics

supplemented by the applicant with additional evidence if the automatic checks produce negative results.While a streamlined and simple process for EU citizens and their families' applying for settled status is important, the use of automated checks must not sacrifice fairness and transparency to apparent efficiency. This use of automated decision making raises some rule of law questions about proper government decision making, including:1) What DWP or HMRC data will be treated as sufficient to be evidence of residence?2) What will the Home Office decision maker be told each type of answers from HMRC/DWP means? What guidance will those decision makers be given on what to do if an applicant's supplementary documents conflict with the results of the automatic checks?3) Will the checks of DWP and HMRC data be integrated, i.e. will the result given to applicants take into account data held by both departments?4) What information will applicants be given on the outcome of the automated checks of DWP and HMRC data by way of reasons for the decision on whether those data show that the applicant has met the residence requirement? 5) How will any errors in the DWP or HMRC data sets that generate wrong results for settled status applicants be identified and addressed?6) What assessment has been made of the risk of data matching between departments not being successful or accurate? What is being done to mitigate the risk of errors in data matching resulting in wrong decisions on the residence requirement?In the first pilot phase, the Home Office has reported that 921 applications were 'checked for automated evidence of residence using HMRC data', of which:• '25 (3%) required some form of intervention to successfully match applicants to HMRC data. Automated matching was not possible primarily due to name matching issues (fixes have been identified and will be applied in a future release) or applicant error (e.g. they entered the NINo incorrectly). Changes to the matching process are being introduced in November to help resolve these issues for future cases.• 13 (1.4%) could not be matched to HMRC records, typically because of data errors, such as NINo and passport records not matching.' It is good that the Home Office is looking into these issues, although these figures suggest issues in 4.4% of cases for the pilot which largely involved people employed by major employers, who are a cohort that is unlikely to encounter issues from the automated checks. It is also unfortunate that the automated checks of DWP data were not trialled in this pilot phase.Furthermore, the issues highlighted by the UN Special Rapporteur concerning how DWP uses HMRC data indicate the need for careful guidance to Home Office decision-makers on the proper assessment of applications. The UN Special Rapporteur explained that, where there were errors in decisions due to the data feed from HMRC to DWP, decision-makers would not pay due regard to evidence that the system was wrong such as pay slips. Instead, decision-makers relied on the incorrect information from the data processing system. The risk of this issue could be mitigated in the guidance given to Home Office decision makers on how to make decisions on settled status. The Home Office could improve its

Dr Cobbe's arguments on the need to uphold a requirement of an explanation for automated government decisions:"Given that whether a public body is obliged to give reasons depends on the circumstances of the case at hand, it has been recognised that reasons may not be required where giving them would be particularly difficult or onerous on the decision-maker. While the argument may be advanced that the opaque nature of automated decision-making systems makes giving reasons onerous or difficult and thus reasons should not be required, this position should be resisted. Rather, a court undertaking judicial review of an automated decision where a requirement to give reasons arises should perhaps consider whether the present inability of automated decision-making systems to provide reasons for a decision should in and of itself be a barrier to the use of these systems for those kinds of decisions in the first place. At a minimum, where the circumstances require reasons but they cannot be provided the court should be entitled to conclude if it wishes to do so that the decision was irrational, provided the facts and circumstances indicate that the decision-making system should have come to a different conclusion (as it would be entitled to conclude if the decision was made by a human). The alternative to these outcomes may result in the use of automated decision-making coming to be seen as a means of escaping accountability." The High-Level Expert Group may also be interested in the work of the AI Now Institute in New York on algorithmic impact assessments, which aims to provide government agencies with 'a practical framework to assess automated decision systems and to ensure public accountability'. The Institute recommends that key elements of such impact assessments include:"Agencies should provide notice to the public disclosing their definition of "automated decision system," existing and proposed systems, and any related self-assessments and researcher review processes before the system has been acquired;Agencies should solicit public comments to clarify concerns and answer outstanding questions; andGovernments should provide enhanced due process mechanisms for affected individuals or communities to challenge inadequate assessments or unfair, biased, or otherwise harmful system uses that agencies have failed to mitigate or correct."

Framework: “Make your work transparent and be accountable”. The Guidance for this principle includes the following:"Your work must be accountable, which is only possible if people are aware of and can understand your work.Being open about your work is critical to helping to make better use of data across government. When discussing your work openly, be transparent about the tools, data, algorithms and the user need (unless there are reasons not to such as fraud or counter-terrorism). Provide your explanations in plain English." However, as discussed by the UN Special Rapporteur, automatic calculation of welfare benefits using HMRC data via the Real Time Information (RTI) system has not been consistent with this principle and guidance. There is a lack of transparency about the automated calculation of welfare benefits through the system, and accountability is undermined by the disregard of evidence that contradicts the automatic calculation. Similarly, the failure of DWP systems to properly address inevitable errors in data from HMRC is inconsistent with principle 4 of the Data Ethics Framework which is “Understand the limitations of the data”. The Guidance for this principle states:"Errors in data are inevitable; however it can be difficult to understand how frequent they are, if they are random, the cause and ways to mitigate or remove them. Errors are not always immediately obvious, especially in large datasets. Simple data visualisations can be the best way of spotting anomalies and systematic errors.You will need to consider and document how identified errors will impact the work.If you find errors in the way data is collected or interpreted, report them to policy or operational staff. "These examples illustrate that guidance and ethical principles are not sufficient – without the rule of law via access to legal remedy, such principles are meaningless.

implementation of the DCMS Data Ethics Framework guidance on transparency and accountability for the settled status process. The Memorandum of Understanding between HMRC and the Home Office for the data access and sharing associated with the settled status process is only available to the public because of HMRC’s response to an FOI request. That Memorandum of Understanding states:"The [application programming interface] API platform provides the ability for Other Government Departments to connect to HMRC APIs. (This content has been withheld because of exemptions in the Freedom of Information Act 2000)Once the OGD has an request registered with the HMRC API Platform, they will be able to use the credentials they have been supplied with (This content has been withheld because of exemptions in the Freedom of Information Act 2000) to access the APIs to which they have been granted access.Home Office will call the HMRC API providing name/NINO and dob. If all three match at the HMRC citizens matching layer then the raw PAYE/SA/Employment data (as detailed in 1. Introduction is sent back to the HO API.Home Office apply their business logic to the raw data which provides an output to the Home Office caseworker of pass/fail/partial pass .The data will not be viewed nor retained by the Home Office. Once the Home Office business logic is applied the Home Office will receive an output of pass/partial pass/fail. Once the output is received the raw data disappears.Even though EU Citizens will make their applications via their personal PC/Laptops they will have no access/linkage to the API. All EU Exit Application API calls are controlled and faciliatated via the Home Office front end platform.The data is expected to be shared between normal business hours and support between 9-5pm. The final support model is currently being developed and the MOU will be updated to reflect the final support details." In terms of transparency and accountability, a number of question arise in response to: ‘Home Office apply their business logic to the raw data which provides an output to the Home Office caseworker of pass/fail/partial pass’. For example, how will this process work? What data will the Home Office business logic treat as sufficient for pass, for fail, and for partial pass? What guidance will Home Office caseworkers be given on how to use the results from automatic checks?From an administrative law perspective, the use of these automated checks of DWP and HMRC in decisions in immigration decisions raises questions about how administrative law and judicial review principles can be fulfilled. How can the question of whether all relevant and no irrelevant considerations were taken into account be answered, noting that the data on which the output of the automatic check is based ‘disappears’ once the output is received.The EU General Data Protection Regulation (GDPR) provides for solely automated decisions under article 22, but there is presently a lacuna in the law regarding decisions that include an automated component. Existing data protections laws on automated decisions will not apply in situations such as the settled status application process.Furthermore, the Home Office has not yet explained how it will address errors in data in relation to the automated checks in the settled status

process, remembering the DCMS Guidance set out above states that 'You will need to consider and document how identified errors will impact the work.'

Aude	Boisseuil	EFFE - European Federation for Family Employment and Homecare	How machines or computer software using artificial intelligence (AI) can behave in an “ethical” way? This is all what the reflexion led by the European Commission is about as well as what has been thoroughly explained by the high level experts’ report. The latter is introduced way in advance so that it can bring matter to the very-likely EU directive on artificial intelligence.To try and make the concept of ethics applied to AI more concrete, the researchers refer themselves to fundamental rights and principles mostly stemming from the Treaty of Maastricht or the Charter of Fundamental Rights of the European Union which both issue major values that are likely to be turned into regulation or coercive rule. In the end those rules should comply with the future norm of a “trustworthy and reliable artificial intelligence”.Civil society actors are capable of taking action at this stage of reflexion.With regard to the complexity of the matter which requires technical and scientific extensive knowledge, our contribution shall bind itself to general suggestions or to stressing out diverse concerns regarding the fundamentals of our sector and its future endeavours.EFFE embodies the sector of family and home employment. It is bound to deliver a political and general reflexion on the evolution of labour and jobs created by citizens at their home (children, disabled and elderly care, domestic tasks...) to the European authorities. For home, place of privacy, of personal, family, social life can also be a place of work for the home employees.	We fully subscribe to fundamental principles that rule this report and place artificial intelligence at the service of human welfare and not as an end in itself. The five ethical principles (do the good, do not harm, protect liberty of choice, be fair, operate in a transparent way) are clear and general enough to allow in every way every possible development of the AI in a near future.	The ten ethical values for a trustworthy and reliable AI happen to be all equally essential from our point of view: - responsibility (1), - data governance (2), - design for all (3), - AI autonomy governance (4), - non-discrimination (5), - respect of human autonomy (6), - respect of privacy (7), - robustness and reliability (8), - security (9), - transparency (10).With regard to our business sector, we shall limit ourselves to focus on only some of those values, on the sole place where it shall occur (private home), and targeted populations (the most vulnerable):Respect of human autonomy (6)Compliance with “respect of human autonomy” (6th value) is a first-rate requirement with regard to the population that is likely to be using those new technologies:Among the first major implementations making a use of the AI at the service of individuals, domestic robots, companion robots, those technologic leaps ask a question to the endeavours of homecare and of social relations. Can companion robots be a response to social isolation? Philosophy and ethics have us considering they can neither replace carers nor professional helpers. They can nonetheless be a considerable discharge of time for humans to execute other tasks in the meantime. A robot can be used for repetitive tasks, without ever replacing the human bond. Our concern will always focus on preserving the irreducibility of human relationship to an automated, robotised or “artificially-altered” one.The notion of vulnerability for individuals (children, disabled, elderly) ought to be particularly significant with regard to the multiples uses that are supposed to improve human welfare out of new implementations facilitating the everyday life. Companion robots can potentially be responsible for “bad treatments” and can also harm “autonomy of decision”, when ethical rules are not taken into account very early;o A robot programmed for his ability to perfectly adapt to his owner could bias some of his purchasing decisions at the expense of his actual financial capacity;o The everyday life with robots may cause long term social risks that must be anticipated. The risk of an individual getting emotionally attached to a companion robot that would simulate empathy, the influence it would have on human actions, its perceptions, consequences in terms of unsocialization are objects that shall be assessed and precisely tested before being massively displayed to the public.o Personal assistants investing our homes and requested to manage multiples tasks to the sound of our voice, among which agenda matters, travels, music (i.e. first use as of today), information research, weather forecasts, home automated equipments are all potentially carrying risk for the autonomy of decision and are likely to exert an influence and bias free will and free process of decision, most particularly for vulnerable segments of population.“Ethics by design” is a requirement that seems particularly adequate to us in the frame of designing and developing social robots at home.Respect of privacy (7)Point 7 (respect of privacy) is questioned by the massive arrival of connected objects such as personal assistants or e-health at home:- The full compliance of systems with GDPR seems difficult to implement, knowing that home is the core center where numerous personal	Through digital transformation, personal housing becomes a connected workplace that is open to exchanges and to the world. Most economic and social flows created nowadays are linked to domestic matters such as e-trade, e-services, e-learning, social media, digital platforms of debate, exchange, social linking and sharing. Personal home hence turns into a marketplace or at least a place of wealth creation (energy, home rentals), economic creation (jobs, remote working, and craft jobs), and social bonding (intergenerational links, education, exchanges, mutual help). It also becomes thanks to technological progress a place of health prevention, of medical care where a significant amount of citizens in France and Europe wish to remain until the end of their life. Artificial intelligence gets massively through to personal homes in function of several different uses made of systems, equipments, software that will be fully implemented in a close future to support people in their daily lives, especially the most vulnerable ones. Lastly, digital transformation is raising hopes regarding the future of caring and housekeeping jobs.EFFE’s reflexion is part of the development of new technologies within a domestic use, driven by the digital revolution and their impact on the employment and skills of citizens, professional and non-professional carers. It is for this reason that we wish to participate in this consultation, making the link between ethics, artificial intelligence, home and skills.	Without going into further detail of the various procedures that will allow the development of the European standard, we fully subscribe to this European approach which tends to differ from the American and Chinese giants by imposing a personal data protective legal framework (GDPR) and likely tomorrow in the field of AI governance. The potential of innovations to come that will be using AI is not intended to replace human intelligence but, on the contrary, to ensure rightful cooperation between human intelligence and computing capacities.For the ethics of a domestic artificial intelligence, we claim a humanist stance. Technological choices are at the service of our political choices and we will always strive to promote an alternative model, respectful of human rights and universal values.

					data can potentially be collected about our lifestyles, our consumption habits, our hygienic behaviour... All of this raises a substantial interest from the Big Four tech companies among which the “domestic personal assistants” that collect personal data and host them outside from European borders.How can we ensure protection of privacy in this context as well as effective implementation of GDPR? All the more that it is known we are under strong incentives to yield our consent for data collection regarding the operators’ leading position.- Home-connected health applications are going to be used more and more;There are many connected objects (heartbeat calculating watch, number of steps, temperature, alarm button to be activated in case of a fall, sensors in the housing connected to the watch made to identify anomalies (too long bed rest, malnutrition in case of refrigerator remaining closed for too long, video and geo location system ...). The ethical question that we want to bring to light is, in the face of "intrusive" sensors in his home, not only the respect of private life (point 7) but also the nature and process of collection of consent of the person vulnerable.→ Regarding sensitive data, this will require great vigilance and transparency on the collection and interpretation of the data, and the resulting solutions or treatments that will be offered to the sick or vulnerable.→ The ethical question remains in the first place the dehumanization of care.The ethical considerations expected in a directive on artificial intelligence will have to mobilize the stakeholders on the evolutions of work and competences, in favor of actions and tasks that will be redefined for skilled jobs so as to be adequately at the service of human relations. Machines shall not replace humans but to alleviate them from repetitive tasks.		
Anonymous	Anonymous	Anonymous	1. Human-centric approach [page i]Fujitsu advocates ‘Human Centric AI’ since 2015. We expect ‘human-centric approach’ in this document should basically coincide with it, and hope we can contribute to this based on our experiences over three years.2. Trustworthy AI made in Europe [page ii]‘Made in Europe’ can be replaced with ‘Implemented/applied in Europe’. If we consider the ‘Made in Europe’ then would it mean the guidelines do not apply to the products that are developed globally or ‘Made outside of Europe’, e.g. USA, Canada, Japan, etc.3. Autonomy of humansReplace ‘Autonomy of humans’ with ‘Autonomy’ as in the AI4People Whitepaper.4. “AI can help humans to identify their biases, and assist them in making less biased decisions.” [page iv]We can put more emphasis on it. This could be a killer application of ‘Trustworthy AI’, which we may not see in other two AI directions in the world – business-first and totalitarianism.	1. The EU’s Rights’ Based Approach to AI Ethics [page 5]Although this is quite valid for the purpose of this document, it would be better to note the relationship between the EU Treaties and Charter of Fundamental Rights, and the Universal Declaration of Human Rights from UN. A ‘rights-based approach to AI ethics’ must be more common in the world, and this approach is an instantiation for the EU.2. “Tensions may arise between the principles when considered from the point of view of an individual compared with the point of view of society, and vice versa.” [page 8]This is a very important message. In this document, it is advised that we should go back to basics – EU Treaties and Charter. However, it is not so easy to do it on a daily basis. We can create our rules and/or bylaws for our specific situations checking if they are fully compatible with the principles (and with priorities).3. “… with equal distribution of economic, social and political opportunity.” [page 9]This is also related to ‘The Principle of Justice’. It is worth mentioning that the principles are NOT mutually exclusive. (Can we say they are collectively exhaustive?)- Freedoms of identifyIs it ‘freedoms of identity’?4. “AI developers and deployers ... they interact with an AI identity.” [page 11]Many AI-based systems are ‘human-in-the-loop’. In these case, do AI developers/deployers need to say ‘AI behind	1. Data Governance [page 14]It is almost impossible to remove all the biases in the data. There are a number of ongoing researches to mitigate biases and to keep fairness not only by eliminating biases in the training data. This document should specify the problems and the objectives (shown in the paragraph 4), but should not specify the methods to address them.2. In large enough data sets these will be diluted since correct actions usually overrun the errors, yet a trace of thereof remains in the data.” [page 15]This statement may not be true because there could be an attack to put subtly polluted data for a long time.3. Design for allIt should be allowed to limit the target users to some extent especially for domestic/local services.4. Robustness [page 17]Currently we do not have appropriate metrics for reliability, reproducibility or resilience to attack (we have some for accuracy). It is an urgent task to establish practical metrics for them.5. Moreover, formal mechanisms are needed to... AI systems. [page 18]Capabilities of current formal mechanisms are too limited to ensure safety of AI systems. It is not appropriate to say ‘needed’.6. This also entails a responsibility for companies to identify from the very beginning of the ethical impact... should comply with. [page 19]The requirement for trustworthiness may vary by application. If EU tries to impose ‘X-by-	1. Reproducibility [page 26]Is it possible to guarantee reproducibility in machine learning systems? It is unlikely.2. What other data sources/models can be used to eliminate bias?It is also mandatory to mention the impact to accuracy when we eliminate bias.3. The Levels of Validation are not set [page 27]For example, in the top bullet point of the ‘Fall-back plan’ the level of validation is unclear in terms of setting the short-mid-long term goals. Also, the criteria for failure are undefined and who can decide if the results are ‘wrong’.4. Safety [page 27]If we entrust the definitions and criteria of safety to the AI developer or AI deployer, then it will not work because the cognitive bias would still be there.	1. Definition of AIThe new definition of AI is very broad and quite ambiguous. If we adhere to the new definition of AI, then each and every offering (system, solution, service) will required to comply with the AI Ethics guidelines and therefore need certification2. ‘Ethics by design’ approach need to be clearly definedThe guidelines strongly suggests the ‘Ethics by design’ approach, however, it is not clearly explained as to how this would be achieved.3. ‘Justice’ and ‘Fairness’ can be treated separatelyIn the interpretation of ethical principles, the authors equate justice and fairness, but there should be a subtle distinction between the two words. More clarity required on the definition of ‘justice’; can we assume that ‘justice’ is equal to ‘be fair’.

				human'?5. LAWS [page 12]LAWS could be more humanitarian than ordinary lethal weapons because they won't cause PTSD for the human soldiers. This topic needs to be discussed on a higher-level including Governments bodies on how to prohibit AI in LAWS.	design' for all the AI 'trustworthy' applications, it may hinder innovation. Flexibility should be introduced.7. Ethical goals and requirements should be integrated at "sense" level ... adherence to those principles. [page 20]Again, it is not appropriate to impose one approach to achieve objectives. Since technologies advance rapidly, there could be different approaches that can make it better.8. In addition, sometimes small changes ... results in dramatic changes in the interpretation, [page 21]It is worth mentioning that sensitivity against data needs to be small enough to avoid unpleasant surprises. Note that it is not advisable to apply machine learning technologies to the system having chaotic behaviours.9. Regulation [page 21]There may be some inconsistencies among current regulations. It is expected that AI HLEG will create well-defined regulation/legislation proposals.	
Jaak	Tepandi	Tallinn University of Technology	AI "made in Europe" may be a good viewpoint from the direct EU perspective. Still AI is developed in cooperation with all countries who reap its benefits, as well as are potentially harmed by its drawbacks. For this reason, the AI "made in Europe" viewpoint seems to diminish potential value of this document and might be less stressed throughout the text.	As AI becomes more integrated with society and human beings, the fundamental rights of human&AI and pure_AI subjects should also be considered. These questions will inevitably appear and it is better to preclude this. Chapter "5. Critical concerns raised by AI", especially section "5.5 Potential longer-term concerns" does not currently present the full extent of well-known critical concerns.	The list of requirements given in Chapter "1. Requirements of Trustworthy AI" seems not to include the idea that the system should do what it is intended to do and only what it is intended to do. To capture this, there might be an additional requirement of (for example) "Minimization of misuse" / "Restricted functionality" or similar. Misuse of AI is a very serious threat. Chapter "2. Non-Technical Methods" might consider the possibility of AI-specific regulations.	The Chapter seems to focus on algorithmic decision-making. Decision-making enabled through other methods, for example using machine learning, may be even more powerful and more problematic as well.	This is a useful document. It might be more influential when extending its viewpoint globally and considering the full positive and negative potential of AI.

Cornelia	Kutterer	Microsoft	Microsoft appreciates the opportunity to comment on the draft Ethics Guidelines for Trustworthy AI ("Guidelines"). As a leading supplier of artificial intelligence ("AI") solutions, Microsoft welcomes the Guidelines' recognition of the tremendous opportunities that AI offers to both individuals and society. We also fully appreciate that new technologies can raise important policy and social challenges, and in this respect AI is no different. Like other technologies that have preceded it, AI will confer enormous benefits on society—but AI systems will also be susceptible to uses that cause harm. We strongly support the efforts of the High-Level Expert Group ("HLEG") to develop a consensus-based framework for the development and deployment of trustworthy AI to mitigate these risks. Microsoft is actively engaged in efforts to develop principles and guidelines in this space, and shared our thoughts in The Future Computed. We have participated in several industry and multi-stakeholder initiatives to support the development and use of trustworthy AI, including the Partnership on AI, the European AI Alliance, the OECD's Expert Group on AI in Society, the 2016 White House consultation on Preparing for the Future of Artificial Intelligence, the Singaporean Government's Advisory Council on the Ethical Use of AI and Data, the AI4People Forum of the Atomium European Institute for Science, Media, and Democracy, the AINow Institute at New York University, the ISO/IEC JTC 1/SC 42 AI standardisation efforts, and many others. Our researchers are also continuing to develop new technologies and mechanisms to address these issues, including through participation in forums such as ACM FAT* and the annual Neural Information Processing Systems conference ("NIPS"). We draw from these initiatives, and our experience in developing and deploying trustworthy technologies more broadly, in our comments below. Many of the positions and recommendations in the Guidelines align with those that Microsoft has articulated and/or publicly supports. For example, we agree that the overall objective of guidance on AI development and deployment should be to maximize benefit while minimizing risk (p. i). We also agree that building trust in AI is fundamental to enabling its broad adoption and realizing its potential (p. i); we understand from long experience that customers will not use technology that they do not trust.We do have a general comment on the scope of the Guidelines, however. We also offer some thoughts on the proposed definitions.1. Scope. The Guidelines provide a thoughtful and comprehensive set of ethical considerations designed to help developers and implementers of AI achieve "trustworthy AI". In offering these considerations, the Guidelines acknowledge that "different situations raise different challenges" (p. 3). We strongly endorse this point, and believe that contextual considerations merit greater attention in the Guidelines. The degree of risk of individual or societal harm, and the potential severity of such harm, will vary enormously depending on the specific AI application at issue. We urge the HLEG to ensure that a careful and thorough assessment of these risks is an integral part of ethical evaluation process set out in the final Guidelines. This risk assessment is vital to ensuring that those who rely on the	Chapter I of the Guidelines identifies core values and principles that those dealing with AI should ascribe to. We broadly agree with the principles, but offer a few further thoughts on structure and content of Chapter I. In terms of structure, the flowchart on p. 4 suggests that the considerations identified in Chapter I are relevant to determining whether the "purpose" of a given AI implementation is ethical; Chapter II in turn identifies "requirements" for AI systems and applications to achieve trustworthiness (e.g., accountability, non-discrimination, transparency etc.). But three of the five factors set out in Chapter I—human agency, fairness, and transparency— seem less relevant to purpose and more relevant to the implementation measures necessary to achieve ethical AI. Indeed, these three criteria are largely repeated in Chapter II. For clarity, we suggest consolidating these three points into Chapter II, and focusing in Chapter I on the purpose of AI and potential use cases that raise particular concerns. Alternatively, dropping references to "purposes" in the context of Chapter I and explaining that these are foundational principles would be helpful.More generally, the HLEG's guidance is rich with ideas and detail. Including annexes or indexes in the final version of the Guidelines, potentially including graphical illustrations of how the various concepts in the Guidelines fit together, would be helpful to make the guidance more user-friendly.We also encourage the HLEG to clarify the ways in which some of the principles are expressed: • Principle of Beneficence. The Principle of Beneficence states that "AI systems should be designed and developed to improve individual and collective wellbeing." While that statement makes sense in the abstract, it is less clear how "improved" wellbeing should be assessed. For example, does the "warehouse storage optimization" example of an AI system provided above meet this criterion? AI can be a tool to improve wellbeing, but it can also serve more neutral objectives whose direct individual or social benefits are less clear. We recommend that the Guidelines adopt a broad understanding of beneficence and acknowledge that AI solutions may satisfy this standard so long as they serve a useful purpose (to someone) that outweighs the risk and severity of potential harm to others.• Principle of non-Maleficence. The Principle of Non-maleficence states that "AI systems should not harm human beings," and should not enhance existing harms or create new ones. Further qualification of that statement would be helpful. Take the case of self-driving cars; if a self-driving vehicle uses an AI solution to choose between hitting one pedestrian or an entire family in another car, does it fail to meet the Principle of Non-maleficence? We encourage the HLEG to more explicitly recognize in this section, and in Chapter I more generally, the fact that there will necessarily need to be a balancing between benefits and harms when deploying AI, and that some trade-offs may be unavoidable. There will be other examples where different parties experience different types of harm and perceive them differently (e.g., an AI solution that helps advertisers more effectively target online advertising versus a solution that helps web surfers more effectively block online ads).	Chapter II identifies "requirements" and "methods" underpinning "trustworthy AI." We welcome the HLEG's thoughtful work to identify these elements, and agree they are key to the process of building trust of individuals and society in AI. We were struck by the repeated use of the word "requirements" in Chapter II. As acknowledged at the start of Chapter II by the HLEG, these "requirements" will always need to account for the context in which AI is deployed. For that reason—and also because the Guidelines are meant to be voluntary—the term "requirement" seems to be a misnomer.In terms of the specific "requirements" and methods addressed in the Guidelines:Requirements• Accountability. We agree that the people who design and deploy AI systems must be accountable for how their systems operate. Indeed, the notion of accountability is central to Microsoft's own articulation of responsible and trusted use of AI. However, we view accountability as part of a broader and more fundamental concept: "responsibility." For instance, developers should be responsible for updating systems in use, if necessary through internal review boards; implementers should understand both the capabilities and limitations of AI systems and take these into account in order to mitigate errors or other harms; users of AI-enabled systems should accept the need to use such systems (such as self-driving cars) responsibly, in line with guidelines and system limitations; and policymakers should seek to understand the impact of changes they propose on AI-powered systems before introducing those changes. This concept is also linked to the importance of human-centric AI. Where appropriate, responsibility also means that the humans that operate using AI systems understand the limitations of AI and are qualified to correct or alter the decisions made by AI. • Data Governance. In some ways, the HLEG conception of "data governance" is too narrow, in that it focuses solely on data. Governance structures necessary to develop AI ethically include a broader range of engineering and design practices as well (for instance, access controls; systems documentation; training for relevant actors; etc.). For example, we refer the HLEG to ISO/IEC 38505-1:2017, which establishes a framework for the governance of data within organizations more broadly, but which could be applicable in the AI development and deployment contexts as well. If we limit our focus only to data governance, we urge the HLEG to recognize that such governance is complex in practice and will need to be tailored to individual scenarios. For example, the Guidelines refer to data retention in order to monitor for malicious inputs to AI datasets. However, that may not always be possible in line with GDPR and other data protection requirements. To take another example, some AI systems may not operate successfully if training or reference data sets are anonymized or deleted, while other systems may essentially require that algorithms are separated from underlying data, in order to be provided through APIs or other services to other third parties consistent with data protection regulations. Governance structures, and data handling practices, therefore must be sensitive to context.• Design for all. The Guidelines' suggest that all AI systems should be	Chapter III sets out an "Assessment List" of questions to help AI developers and deployers to assess the trustworthiness of AI pursuant to the principles described in the Guidelines. We support the HLEG's efforts to provide this sort of concrete guidance, which is important to enable innovators to understand in a practical way how ethical principles can be embodied in products and services. We also agree with many of the questions proposed. We can also anticipate additional questions that may be relevant, depending on the use case. For example: • Governing AI autonomy: (If applicable) Is there a mechanism in place to allow affected individuals to request, and receive in a timely manner, human review and revisiting of consequential decisions made by AI systems? • Robustness: How does the system handle unexpected events or unexpected interactions with individuals? • Purpose: Is there a clear purpose for developing or deploying the AI system? Is the purpose an ethical one?In terms of how the questions are applied, we agree strongly with the statement in the draft Guidelines that the precise questions that are relevant to assessment of any particular AI system will vary depending on the use case. Development and deployment of AI systems in the healthcare sector provides a good example of why tailoring of the Assessment List will be required depending on context: • Health applications are already subject to a well-developed regulatory regime for their development and use, including standards for ensuring that the benefits outweigh the risks and that technologies are safe and reliable. While evaluating AI-based technologies under these existing regulatory frameworks may present some challenges, any assessment of medical AI systems will need to reflect these existing regulatory assessments (e.g., safety and effectiveness) and requirements. Further, AI systems may also be deployed in healthcare settings such as clinical trials that require prioritizing different considerations than the principles of the Guidelines. For example, in a blinded, randomized clinical trial to assess a AI-based technology, in order to maintain the blinded nature of the study, the ability of the developers to provide transparency around the technology may be more limited than when deployed outside the clinical trial setting. • One particular challenge for AI-based medical technologies is ensuring that the output is not just technically accurate (that the correlation or rule that an AI system learns accurately reflects the data) but also clinically reliable. Determining whether a correlation or rule implemented by an AI system properly characterizes clinically relevant variables and cause and effect requires unique expertise. Thus, when assessing the "accuracy" of AI systems in healthcare use cases, it is particularly important that domain experts be involved in the development and assessment. Further, this challenge reinforces the need for clinical reasoning and judgment in the use of AI systems.• Because decisions made by AI systems in the health domain have the potential to impact patients' health and care, we believe it is particularly important that those deploying and relying on these system (e.g., healthcare professionals, patients, managed care organizations, regulators) understand how the systems make decisions. This requires not just an

Guidelines appropriately evaluate which criteria set out in the Guidelines apply and, if so, how they should apply. Some implementations of AI will have nominal or inconsequential impacts on individuals or society. Take, for example, the case of an AI system designed to optimize storage of items in a warehouse. In such cases, several of the recommendations in the Guidelines—such as "the presence of an internal and external (ethical) expert . . . to accompany the design, development and deployment of AI" (p. 8)—might be inappropriate, or even nonsensical. In fact, many of the ethical issues identified in these Guidelines only arise for AI systems that have a consequential—or meaningful—impact on individuals. We encourage the HLEG to make clear at the outset of the Guidelines that their recommendations are not "one-size-fits-all" and instead should be tailored to each specific implementation of AI depending on a careful and thorough risk assessment. Providing a framework to assist those deploying AI in conducting this risk assessment (e.g., perhaps drawing upon learnings from ISO/IEC Technical Report 27013:2018, which provides guidance on leveraging existing standards in a cybersecurity risk assessment framework) would be useful.2. Definitions. • "AI." The Guidelines define the term "AI" as systems designed by humans that perceive their environment through interpretation of data, reason on the basis of that data, and "decid[e] the best action(s) to take (according to pre-defined parameters) to achieve the given goal." (p. iii). We offer two observations on this definition.First, it would be helpful for the final Guidelines to provide greater clarity on what is meant by "deciding the best action(s) to take." If an AI solution ranks pieces of information (e.g., book titles, search results, names, etc.) based on various inputs, is it "deciding" an "action to take"? If an AI solution uses data inputs to "score" employees based on their likelihood of taking sick days, but leaves it to human actors to decide how to interpret or apply the score, is the AI "deciding the best action(s) to take"? Article 22 of the GDPR articulates the concept of a "decision based solely on automated processing"; is the AI definition set forth in the Guidelines coextensive with the GDPR concept of solely automated decision-making, or it is narrower (or broader)? We also note that the Guidelines' definition of AI is narrower than many common understandings of the term. As Eric Horvitz, director of Microsoft Research Labs, has noted, the term AI is often used to mean "a set of computer science disciplines aimed at the scientific understanding of the mechanisms underlying thought and intelligent behavior and the embodiment of these principles in machines that can deliver value to people and society." The draft Guidelines, by contrast, define AI as systems that "decid[e] the best action(s) to take." Many solutions in use today that are described as having an AI component do not necessarily "decide" on a course of action; instead, they make connections, reveal correlations, or provide other insights that humans then use to decide on a course of action. If the HLEG chooses to define AI as systems that "decide" on a course of action, the Guidelines should acknowledge that many solutions that are commonly understood to constitute or incorporate AI

Advancing the interests of certain individuals may inevitably impose harms on others (e.g., an AI tool that makes one company more efficient might "harm" rivals by making them relatively less able to compete). We suspect that the HLEG does not intend this Principle's prohibition on harm to forbid these scenarios, and greater clarity on this point in the final Guidelines would be welcome. These points also reinforce the importance of the accountability principle in ensuring that human actors ultimately remain accountable for the operation of any AI system, including how they balance potential benefits and harms.• Principle of Fairness. The Principle of Fairness provides that AI developers and implementers "must ensure" that individuals and minority groups remain free from bias, stigmatization and discrimination. Microsoft strongly supports the view that AI should never be used to engage in unlawful discrimination, and that all parties involved in AI development and deployment should commit to mitigating AI outcomes that impose unfair biases (see, for instance, our Six Principles for Developing and Deploying Facial Recognition Technologies, cited above). Decisions made at every stage of AI development and deployment can inadvertently inject bias. Efforts to remove unfair bias from AI systems—similar to efforts to promote privacy and security—should be considered at every stage, starting with task definition and continuing all the way to system deployment and feedback. As noted in Part I of this response, however, removing all forms of bias from any AI system or finite dataset might not be possible, which the draft Guidelines themselves recognize (see p. 16). Thus, we would encourage the HLEG to revise this Principle to focus on addressing "unfair" biases in AI systems. For instance, the Guidelines could encourage developers of AI to disclose, in appropriate cases, key features and limitations of the datasets on which the AI was trained, and for AI implementers to take these limitations into account in order to mitigate the risk that the AI might generate unfair outcomes for specific individuals or groups.Critical concernsThe HLEG asks specifically for stakeholder input on Section 5, the "critical concerns" raised by AI.In general, we believe that the HLEG has taken the right approach in choosing to cast these as "concerns," rather than as "red lines." As the Guidelines state, a balance is required between what can be done with AI and what should be done with AI. In some scenarios, the concerns identified in this section might lead to the decision not to implement a particular AI solution. But in others, these concerns might be mitigated in other ways, or might be outweighed by other interests. In the case of an AI system that allows for normative scoring, for example, the inclusion of a mechanism that enables human review and correction can mitigate potential risks of harm and enable ethical deployment of the system.In terms of the concerns themselves, we offer the following thoughts.Identification without consentAlthough we agree that identification without consent could be a critical concern in some scenarios, it might not be a critical concern in others. In some uniquely sensitive AI implementations—e.g., certain uses of facial recognition technology to uniquely identify individuals—providing a robust notice-and-consent experience to

designed for use by all categories of people in all cases. This statement strikes us as overly broad, and we suspect is not what the HLEG intended. Should, for instance, an AI system designed to aid lorry drivers be made accessible to all ages—even children or individuals whose vision is not sufficient for a legal driving license? Instead, the key in our view is to design systems that enhance accessibility where possible, and that are accessible for all persons in similar situations, including those with disabilities or in minority groups. AI technologies hold tremendous possibilities for people with disabilities. Microsoft strongly supports accessibility both in, and assisted by, AI, and has made significant investments to develop AI to amplify human capabilities. For example, the "Seeing AI" app, which Microsoft makes freely available, seeks to narrate the input into smartphone cameras in order to benefit the low vision community. Microsoft has also developed a pictogram app, Helpicto, to help children who are nonverbal to communicate. • Robustness. Microsoft agrees that the robustness of each AI system is a key consideration—but it is also important to recognize that the "perfect should not be the enemy of the good." In addition, we encourage the HLEG to consider the "robustness" of AI systems relative to today's—often even more error-prone—status quo. That is not to say that AI developers should not strive for perfection, and other elements of robust systems, such as reproducibility. Indeed, certain Microsoft AI systems, including Azure Machine Learning Services, are able to store training data models and document them in order to help aid reproducibility.• Governance of AI Autonomy. Microsoft strongly believes that humans—not machines—should remain "at the center" of the system and ultimately responsible for AI. A corollary of this is that AI should not be designed to replace humans; rather it should be designed as a tool to enhance and expand their capabilities—i.e. "augmented," rather than artificial, intelligence. We also agree with the HLEG that a "human-centric" approach can, depending on the scenario, require special efforts to explain system outputs to humans affected by the system. Such an approach can also entail levels of human input and control, up to and including (depending on context) scenarios where humans can step in and alter decisions where errors can be clearly recognized or corrected. To achieve this, we have called for laws requiring parties that deploy facial recognition to undertake meaningful human review of facial recognition results prior to making final decisions for what the law deems to be "consequential use cases" that affect consumers. This includes where decisions may create a risk of bodily or emotional harm to a consumer, where there may be implications on human or fundamental rights, or where a consumer's personal freedom or privacy may be impinged.• Transparency. In particular when AI is used to help make decisions that impact people's lives, we agree it is critically important that people understand how those decisions are made. This covers, under the broad umbrella of "transparency," a host of important aims, including both "explicability" (as discussed in the Guidelines) and what we have termed "intelligibility" (simply put, useful explanations of the behavior of AI

explanation of how the AI system produces its results but also contextual information about how the system works and interacts with data to enable the medical community to identify and raise awareness of potential bias, errors and other unintended outcomes. In addition, affected individuals should understand clearly the intended role of the system in medical decision-making. If healthcare professionals do not understand the limitations of AI systems (including accuracy) or misunderstand the role of the system's output, unfairness may result. • However, given the scale of the data and computational processing utilized in these technologies, it may not be possible for a clinician to understand the data analyzed by the technology, let alone understand how the data results in the recommendation offered by the technology. This will make it more difficult for the healthcare community to assess whether the technologies propagate biases or underrepresentation inherent in the underlying dataset, and whether the technology is clinically accurate in addition to being technically accurate. In light of these challenges, it is particularly important that when AI is deployed for healthcare, it should augment the skills and experience of clinicians, rather than replace those skills.

will fall outside this definition, and therefore outside the Guidelines' scope. Ultimately, to avoid confusion over the Guidelines' scope and application, we encourage the HLEG to more clearly define AI, and to ensure that all of the scenarios and illustrative use cases set forth in the final Guidelines fall within the scope of the final definition. • "Bias." The Guidelines define "bias" as "prejudice for or against something or somebody, that may result in unfair decisions" (p. iv). In the views of most data scientists, virtually any dataset will reflect at least some types of bias (e.g., traffic data collected in large cities might not accurately reflect traffic patterns in smaller cities; data about social media use by teenagers might not accurately reflect usage patterns by older users; etc.). The goal should not be to eliminate all biases in datasets used to train AI, as this is effectively impossible for most (and possibly all) finite datasets. Rather, the goals should be: (1) to help people understand the scope, characteristics, and limitations of the dataset(s) on which an AI solution was trained, so that people can better understand how these limitations might impact the outputs generated by the AI in any given application; and (2) to ensure, to the extent possible, that AI systems do not result in harms associated with undesirable human biases. Indeed, the Guidelines appear to recognize this point on p. 16 ("While it might be possible to remove clearly identified and unwanted bias when collecting data, data always carriers some kind of bias."). We urge the final Guidelines to revise the definition of "bias" to more clearly reflect these points.• "Ethical purpose." The Guidelines define "ethical purpose" to mean AI that "ensures compliance with fundamental rights . . ." (p. v). This equation of ethical purpose with compliance with fundamental rights—and in particular with rights set out in the EU Treaties and the EU Charter of Fundamental Rights—is carried through throughout the Guidelines (see. e.g., p. 3). While the aspiration to comply with fundamental rights is critically important, the final Guidelines should also recognize that the nature of obligations flowing from these rights could vary significantly depending on the context, and thus might not always be apparent to those developing AI systems. For instance, with regard to freedom of expression, there might be a fundamental right interest in permitting a private-sector actor to utilize AI for certain purposes (e.g., to automatically identify and post content on a website only if it reflects a particular political viewpoint), where the same AI application adopted by a public-sector actor might infringe upon fundamental rights. We would encourage the final Guidelines to note that the application of the obligation to respect fundamental rights could vary significantly depending on the affected individual at issue and specific AI application at issue.• "Users" and "developers". Although the Guidelines do not define the term "user" or "developer," we urge the HLEG to give these terms further thought and to provide greater clarity on their meaning and consistency in their use. On "users", the Guidelines at times employ the term to mean the person who is making use of the AI (e.g., the bank officer who uses AI to "score" an applicant for a loan) and at other times to mean a person who is impacted by the AI (e.g., the loan

individuals might be warranted. In other scenarios, however, identification without consent might be justified less concerning—or even more compelling. We recommend that the final Guidelines' discussion of this issue focus specifically on use cases where identification without consent posed an elevated risk of harm to individuals or society. We also recommend that the Guidelines expressly acknowledge that different applications of AI might warrant different types of consent. In higher risk scenarios, explicit consent might be appropriate, while in lower-risk scenarios, consent may be expressed implicitly, e.g., by clearly informing a consumer that stepping into a store will entail the use of AI tracking to enable "frictionless" shopping experiences. In addition, the final Guidelines should note that many of these issues relating to identification—and so to processing of personal data—are already governed by the GDPR and other EU law. It is not clear at this time that further requirements are needed simply because the data processing is carried out in combination with AI technologies. This discussion about consent raises a more general and fundamental point about consent and AI ethics. We note that, at times, the Guidelines recommend opt-outs or informed consent as a pre-requisite for ethical AI. For instance, the Guidelines suggest that informed consent is necessary to respect the Principle of Non-maleficence (i.e., do no harm). But this conflates the processing of personal data—an activity that is intrinsically linked to individuals by definition—with AI technologies. AI may or may not rely on personal data processing, and may or may not affect individuals. In the case of many AI applications, the consent of the individual to the use of the AI will be unrelated to the "ethics" or "trustworthiness" of the system. It may also be unclear which individuals should consent. Take, for example, an AI system in an autonomous vehicle designed to avoid accidents; society might have an interest in ensuring that the AI system is engaged (and thus traffic accidents are avoided) that overrides a passenger's refusal to "consent" to use of the AI. That is not to say consent is not important—in our view, informed consent is a key tool to give data subjects control over how their personal information is used in AI applications. But the issues consent raises are more complex than the Guidelines currently suggest.Covert AI systemsThe Guidelines state that "[a] human always has to know if she/he is interacting with a human being or a machine, and it is the responsibility of AI developers and deployers that this is reliably achieved" (p. 11). We would note that principle is potentially under- and over-inclusive, depending on how one understands the notion of "interacting." For instance, search engines often use AI to rank results, but sometimes a user will be presented with results that have been ranked in part by humans (typically to help test and improve the search ranking algorithm). Requiring search engine operators to notify users in all cases seems unnecessary and could actually undermine the usefulness of such testing. On the other hand, a person might be significantly impacted by an AI without "interacting" with it (e.g., where a judge uses AI to help determine a criminal defendant's prison

systems and their components). Achieving this aim in practice can be complex and highly dependent on a host of variables, precluding anything resembling a "one-size-fits-all" approach. Certain AI technologies, including deep neural networks, typically involve thousands of parameters and incredibly complex interactions between input features that go well beyond what is comprehensible to humans. This is their strength; their complexity enables them to more accurately solve problems in challenging domains like computer vision and natural language processing. It also means, however, that people cannot understand how they work by simply observing their internals. In these cases, the overall goal of "transparency" would be ill-served by attempts to examine the structure or parameters of more complex models or the source code used to implement them, due to the complexity of that information and its irrelevance to a practical description of their behavior. Instead, more considered solutions that are better tailored to the audience—humans—are required.A number of promising technical approaches to achieving intelligibility of both system components, including data and individual models, as well as entire systems have begun to emerge, and we recommend that the Guidelines reference them. They include:o Illuminating datasets with "datasheets." A group of researchers at Microsoft has recently initiated a project named "datasheets for datasets." The project replicates the common practice in the electronics industry of accompanying every component, no matter how simple, with a datasheet detailing standard operating characteristics, test results, recommended usage, and other information. The datasheets for datasets project similarly recommends that every dataset used for AI training be accompanied by a datasheet that describes and explains its motivations, its composition, how it was collected and pre-processed, and any limitations in the dataset that could result in unintended outcomes, such as known biases. Work has also started to develop similar datasheets for documenting critical information about models and systems. o Local Interpretable Model-Agnostic Explanations (LIME). These explanations for individual outputs or predictions work by learning a simple model that approximates the behavior of the underlying model or system for each such output or prediction to be explained. In the context of image classification, the simple model might help a developer understand why a system has incorrectly labeled an image of a husky as a wolf by revealing that this prediction was likely influenced by the presence of a snowy background, which the system commonly observed in training images of wolves.o Counterfactual Explanations. These generate explanations for individual outputs or predictions by identifying how changes to inputs would cause the model or system to produce a desired output or prediction.o Black-Box Explanations through Transparent Approximations. These focus on explaining the overall behavior of a model or system by using outputs or predictions to learn a few sets of simple decision rules, each of which offers an explanation for the behavior of the underlying model or system for a particular range of input feature

applicant). This inconsistency in usage could cause confusion and should be addressed in the final Guidelines. Likewise, the Guidelines sometimes use the term "developer" to mean the natural person who develops the AI at issue, and at other times to mean the entity that offers the AI to potential customers. Here again, greater clarity and consistency in the use of this term would be helpful.• "Transparency," "explicability" and "explainability". The Guidelines frequently use of each of these three terms, often apparently interchangeably. In our view, "transparency" is a broader concept than "explicability." In addition, the latter term is also linked to an important separate term, "intelligibility," that is somewhat overlooked in the Guidelines. We would encourage the HLEG to include each of these four terms in the glossary to help clarify intended meanings for stakeholders, and to use each one consistently throughout the document. Given the critical nature of transparency in the ethical use of AI, more focus on these concepts in the final Guidelines is imperative.In addition, we note that the Guidelines frequently use words such as "requirement" and "compliance." These terms could be (mis)read to suggest that adherence to the Guidelines is mandatory or subject to formal conformity assessment requirements (which, as the Guidelines note, is not the case—see p. i), or that adherence to the Guidelines is the only way to achieve ethical AI. There are a variety of ways to achieve ethical AI, and how best to do so may vary depending on the nature of the AI application in question, the voluntary governance practices of the organization or individual(s) responsible for the creation of the AI, and the use to which the AI is put. We encourage the HLEG to make this point clearer in the final Guidelines, including by consider terms other than "requirement" and "compliance" in appropriate instances throughout the document.Finally, while we appreciate the HLEG is a European endeavor, AI is being developed in a global context. Policymakers and stakeholders in jurisdictions outside the EU might take a different approach to the issue of "ethical AI" based on their unique cultures and contexts, and AI developers and implementers might need to take account of many different approaches at once. For instance, some cultures balance the interests of individuals versus society differently than in the EU; others might place a different priority on animal rights or environmental concerns. In addition, AI solutions may be composed of multiple elements where the machine learning that produces a model occurs in one jurisdiction by one firm, and the AI system that uses that model is built by a different firm in a different jurisdiction. In the section entitled Scope of the Guidelines (page 3), the HLEG asserts that the Guidelines assume that "AI developers, deployers and users [will] comply with fundamental rights and with all applicable regulations." Yet if the development and training of an AI system and the corresponding governance practices happen outside of the EU, but the deployment and use are within the EU, the concept of what rights and applicable regulations apply will need to be understood in a global context. We urge the HLEG to take account of this fact by striving to make the final Guidelines interoperable with other good governance efforts to the greatest sentence). Although "covert" might not be the best term to describe any of these situations, these examples suggest that a better approach to this issue might be to make notification turn on the degree to which the use of AI might have a consequential, negative impact on a person. Potential longer-term concerns We also note the HLEG's consideration of "black swan" technologies such as general artificial intelligence. We agree that such technologies certainly may raise novel challenges, many of which we cannot anticipate today. The goal of the Guidelines, and of any ethical principles in this space more generally, should be technology neutrality. We should aim to develop principles that are sufficiently flexible and enduring to address future challenges created by yet-to-be-developed technologies as and when they arise.

combinations.Technical Methods to Achieve Trustworthy AIThe methods described in this part of the Guidelines (pp. 19-21) are welcome and will often be relevant to the development of trusted AI. To avoid fragmentation and to further expand a common understanding of methodologies used, the methods in this section should be expanded to account for existing best practice and standards (including, e.g., ISO/IEC JTC 1/SC 42). • Auditability. We recognize the need for auditability of AI where impacts are potentially significant. For example, we have endorsed the principle that providers of commercial facial recognition services enable third parties engaged in independent testing to conduct and publish reasonable tests of their facial recognition services for accuracy and unfair bias. However, the nature of auditability will be heavily context-dependent. In some cases—for instance where AI is used to navigate commercial aircraft—regulators may need to be able to audit and understand the details of AI decision-making. In other cases, third party auditors and expert reports will be more effective (given their ability to specialize and understand technical detail). In still other scenarios, internal organizational auditing and controls may suffice. The Guidelines should do more to acknowledge that effective auditing, depending on the context, can include any of these mechanisms.• Risk-based approach. Although the draft Guidelines do not discuss risk management as a tool to achieve Trustworthy AI, we would encourage the HLEG to add a discussion of this point in the final Guidelines. In our view, risk management is at the core of advancing trustworthy AI. Those who develop or implement AI should be responsible for conducting appropriate, robust risk assessments, and for mitigating identified risks through effective safeguards and controls, such that the benefits of the AI implementation outweigh any residual risks. This is fundamental to ensuring that users and other stakeholders are protected. A consistent, risk-based, outcome-focused approach can help businesses develop the trust of their customers and society by enabling them to demonstrate the controls in place to protect users and others from harm. Microsoft is a strong supporter of efforts to create risk-based assessment regimes, as these can help organizations craft effective solutions to significant technical and operational issues in new technologies. Examples of such frameworks include ISO/IEC Technical Report 27013:2018 and the NIST Cybersecurity Framework. HLEG should consider advancing the use of risk assessments for AI as a tool for companies to interpret how technological, operational, and policy controls, requirements, and standards can support implementation of Trustworthy AI. The final Guidance should also embody a risk-based approach to assist companies in deciding where they should most effectively focus their efforts. As the references in the preceding paragraph indicate, regulators, standards organizations, and industry have long recognized the benefits of a risk-based approach in the related context of cybersecurity. As NIST notes in its Questions and Answers on the Cybersecurity Framework, it is "not a one-size-fits-all approach [to managing cybersecurity risk] for all critical

extent possible, including engaging and conducting dialogues with relevant organizations and stakeholders outside the EU and supporting multilateral mechanisms such as international standardization to achieve coherence across jurisdictions.

infrastructure" because such organizations "will continue to have unique risks—different threats, different vulnerabilities, different risk tolerances—and how they implement the practices in the Framework to achieve positive outcomes will vary." Moreover, the Cybersecurity Framework maturity model recognizes that an "adaptive" approach to cybersecurity risk management is the most sophisticated. Similarly, in the financial institutions context, the U.S. Federal Financial Institutions Examination Council has issued a Cybersecurity Assessment Tool, based largely on the NIST Cybersecurity Framework, to assist organizations in determining the relationship between their inherent risk and readiness to address that risk. As with the NIST Cybersecurity Framework, an AI risk assessment model should enable businesses to assess their current AI compliance practices and their compliance goals, based on an informed perspective of the relevant risks. That approach allows businesses to innovate and to determine the appropriate solution for their products and services in a way that is flexible and context-aware. In contrast, an inventory of prescribed controls would necessarily adopt a "lower common denominator" approach and render AI guidance useless for any organization that does not fit the predefined mold. It would risk transforming the document into a compliance checklist, rather than a dynamic tool for identifying and managing risk. Like the frameworks discussed above, any risk assessment approach set out in the final Guidelines also should be outcomes-focused (see, for example, http://ethicstoolkit.ai). http://ethicstoolkit.ai/ Non-Technical Methods to Achieve Trustworthy AIWe welcome and agree with the HLEG's identified non-technical methods for ensuring an ethical and robust AI. Indeed, many of these methods, such as standards, or stakeholder dialogues, are already either in progress or development. RegulationMicrosoft welcomes the HLEG's plans to review applicable regulation. In doing so, we encourage it to consider the extent to which existing law already achieves some of the HLEG's goals. In particular, more thought needs to be given as to how the HLEG principles and the GDPR intersect and will co-exist. There is meaningful overlap between the obligations in the GDPR and the requirements for "Trustworthy AI" as set out in Chapter II of the Guidelines. For example, the GDPR already imposes obligations on data controllers and data processors handling personal data in relation to transparency, robustness, accountability and data governance, and also includes specific restrictions on automated decision-making with significant effects. Importantly, the objectives of the GDPR are not confined to data protection. Instead, as the GDPR explains, the Regulation is "intended to contribute to the accomplishment of an area of freedom, security and justice and of an economic union, to economic and social progress, . . . and to the well-being of natural persons" (GDPR Recital 2). These aims flow down into specific obligations in the GDPR, such as the requirement to provide human intervention in cases of automated decision-making with significant effects, the fairness principle, and the obligation to carry out data protection impact assessments ("DPIAs") for processing

likely to result in a high risk to the rights and freedoms of individuals. In regard to DPIAs, for example, the EDPB has indicated that the reference to "the rights and freedoms" of data subjects "primarily concerns the rights to data protection and privacy but may also involve other fundamental rights such as freedom of speech, freedom of thought, freedom of movement, prohibition of discrimination, right to liberty, conscience and religion." (emphasis added). The overlap between the GDPR's obligations and the HLEG principles does not rule out the need for further regulation in certain areas, of course. As described above, for example, we can see the need for new legislation in relation to certain aspects of facial recognition technology. More generally, as HLEG reviews the EU regulatory existing framework affecting AI beyond GDPR, we also encourage it to consider both where new measures are needed, and where existing legal requirements, or lack of clarity in what the law requires, may pose unnecessary impediments to AI innovation. We also would urge the HLEG to seek to identify areas where regulatory sandboxes can be effectively and usefully deployed. Given the many benefits that AI can deliver, we hope that the HLEG will review the current legal framework with both aims—ensuring ethical AI, and also enabling ethical AI—in mind.StandardizationMicrosoft agrees that standardization has an important role to play in relation to achieving trustworthy AI. International standards can establish coherent and consistent understanding of foundational concepts, management and governance practices that incorporate disciplines from privacy, cybersecurity, reliability and product safety. Standardization also acts as a mechanism for sector-specific and application-specific objectives to be clarified in a manner that is both actionable and flexible for industry. In addition, the use of international standards supports innovation and market opportunity to the net benefit of all participants in the international community while being broadly applicable without prejudice to cultural norms. We also recommend, however, that the final Guidelines clarify that standards should not be used as a replacement for the development or updating of laws or regulations related to AI. Diversity and inclusive design teamsThe Guidelines state that it is "critical that . . . the teams that design, develop, test, and maintain [AI] systems reflect the diversity of users and of society in general" (p. 22). We agree with this point, and have articulated similar principles and commitments both in our Principles on Facial Recognition Technology and in our Guidelines on the development of conversational AI (chatbots). We think such diversity is particularly important in the design and development of AI systems, and we support calls for developers to strive for greater diversity amongst coding teams to help ensure that AI systems operate as fairly as possible.

Erik	O DONOVAN	Ibec (EU Transparency Register ID: 479468313744-50)	Ibec: • Thank the AI HLEG for their important work and the opportunity to comment on the draft guidelines. The adoption of AI technologies is relevant to further economic opportunity, competitiveness and well-being across Europe. • Support the need for a human-centric approach to AI and agree that 'trustworthy AI' has two components – an 'ethical purpose' and being 'technically robust'. • Welcome the recognition that a 'tailored approach is needed [to the implementation of the guidelines] given AI's context specificity'. • Welcome the AI HLEG intention to: o Develop not only a set of ethical values and principles, but to offer guidance on the implementation of these in the development and use of AI systems. o View the guidelines as a 'living document', capable of accommodating new insights and market developments in AI. On this point, the Section entitled, 'Role of AI Ethics' (Page 2) states that the document is 'the beginning' of a process of discussion. This should be clarified in the document – (a) how the implied process of stakeholder engagement/discussion would take place and (b) how the document should be used as a reference in the development of national AI strategies - aiding harmonised implementation across the EU Digital Single Market. o Foster reflection and discussion at global level – the development of international frameworks on AI is hugely important. Perhaps the ambition could go further in fostering reflection, discussion and agreement at global level?	• Freedom of the individual (3.2, page 7) This section could clarify how the freedom of an individual is balanced with the freedom of other individuals and national security obligations. • We broadly support the five ethical principles suggested (do good, do no harm, preserve human agency, be fair and operated transparently). Some additional comments on the principles: o Section 4 (AI context and correlating values, page 8) notes that tensions may arise between ethical principles as viewed by an individual versus as viewed by society. Firms would welcome further guidance on this point – is there hierarchy that needs to be applied when faced by a contradiction between the principles? o On 'beneficence' (page 8), does the draft's statement, 'AI systems should be designed and developed to improve individual and collective well-being' interfere with the conduct of contracts by private actors? o On 'non-maleficence' (page 9), the qualifying concept of intent should be introduced e.g. 'AI systems should not be designed in a way that intentionally enhances existing harms or creates new harms'. o On 'autonomy' (page 9), individuals should have the right to know that they are interacting with an AI or not. However, the right to opt out should depend on the context. The right to opt-out could conflict with other legal obligations. o On 'justice' (page 10) is the requirement that '…negatives resulting from AI should be evenly distributed' realistic? Would it not be better to mitigate or prevent negatives than share them? o On 'explicability' (page 10), the first requirement should be to inform individuals that they are interacting with an AI system, the basis of transparency. Informed consent is important but under GDPR legitimate interest is also allowed for data processing.	• We broadly support the ten requirements of Trustworthy AI (pages 14-28). However, the document could benefit from merging some requirements and making some slight adjustments to make the requirements easier to follow and implement e.g. o 'data governance' and 'respect for privacy' o 'design for all' and 'non-discrimination' o 'governance of AI Autonomy' and 'respect for human autonomy' o 'robustness' and 'safety' • Some additional comments on the ten requirements of Trustworthy AI (pages 14-28): o On 'data governance', in certain cases it should be recognised that a certain bias is intended because of the objective of the application. This should be elaborated on – is there a risk of over deletion? Again, it is acknowledged that the objective must be to ensure that bias does not lead to unfair discrimination. o On 'design for all', should clarify that this relates to accessibility. Is it possible that every system can designed in such a way that it can be used by all? o On 'robustness', more guidance would be welcome on the level of accuracy required for AI systems, particularly sensitive use cases. The use of 'fall-back plans' may be dependent on the use case. o We support the requirement of 'transparency'. • The draft provides a positive non-exhaustive list of technical and non-technical methods to achieve trustworthy AI. However, the use of standardisation and codes of conduct could be emphasised more e.g. standards could be linked to the 'ethics by design'.	• It is acknowledged that the assessment list provided is preliminary and requires more work. The envisaged approach of continuous assessment/improvement is interesting. The questions will need to be clearer and more detailed in order to be accessible, instructive and implementable. Suggested answers, dependent on use cases, would also be welcome. This will help developers and firms in implementation. • The impact of ethical purpose on innovation should be assessed. • The scope of the requirements and application of this guidance should be contextual and risk-based. For example, not every requirement may be applicable to every firm or context. Clarity on this would be welcome.	• The point on sign-up should be clarified further. What will be involved in a sign-up to the guidelines?
Anonymous	Anonymous	Anonymous						Thanks for the guidelines. These are helpful and a good advance.A few general comments on my end:- there is not much talk about the liability problem (i.e. when an algorithm self-develops and provides services to a customer: who is responsible? the creator? even if the creator has been following guidelines the algorithm may change into something that causes harm (say financial) to the user- what about AI that also designs AI? how do we govern that piece?- the role of ethics: my belief is that ethics should be a guiding principle for design, but it should not necessarily be a goal for AI. AI would probably have a commercial goal and as long as it abides to the principles engrained (i.e. these guidelines or, in science fiction, the rules of asimov with robots) we should be ok in this regard.- I miss something more about super AI and singularity: I understand that probably it is part of section 5.5. in page 12 explanation but it would have been good maybe to mention that the ethical principles should be engrained in a way that super AI would not be able to override. But again, this is probably a lot of speculation on my part.- maybe it would be useful to have a bit more explanation regarding principles guiding AI that may integrate with humans (somehow, in the words of Mr. Musk, we are already

								cyborgs with our mobile phones).In any case, I would like to thank the group for working on this matter and being forward thinking. I really enjoyed reading through the document and appreciate the mindset the expert group has.
Kristin	STRAUCH	Bitkom (Bundesverband Informationswirtschaft, Telekommunikation und neue Medien)	As Europe's largest digital association, Bitkom endorses the published draft of the AI Ethics Guidelines and welcomes that the issue is being addressed in an interdisciplinary manner at the European level. We especially support the Commission‘s engagement to interact with a broader set of stakeholders in order to develop these guidelines and to share information on the Group's and the Commission's work. The paper represents an important and valuable approach to specify how concrete ethical values can be operationalised in the social, political and economic context of AI. Europe’s ethical values should not only be implemented in the development of AI but also facilitate socioeconomic progress. As such, digital ethics can represent a significant competitive advantage within the field of AI. Bitkom wishes to emphasize that ethical guidelines should be sharply separated from legal issues. Our understanding is that the guidelines are intended to contribute to lever-aging this potential. We understand that the guidelines neither constitute nor directly prepare new regulation regarding AI. A tightened legal framework would be detrimental to the European AI ecosystem and thereby constitute a societal disadvantage.	At numerous places, terms like „wellbeing“ and „the common good“ (p. 5) are being addressed without further interpretation. In order to avoid uncertainty, we suggest that the authors should ensure that these terms relate to civil rights. Moreover, we would like to note that the section “Fundamental Rights of Human Beings” (p. 6f) does not address the important questions about value assessment and value conflicts. The paper stresses that in case of harm, AI systems should provide users “with effective redress” (p. 10). The question of users and AI system operators responsibility should be further clarified. Ultimately, a diffusion of responsibilities could trigger distrust among users and thus stifle the overall acceptance of AI systems. In our view, the guidelines require further concretization regarding the monitoring and willingness to assume responsibility of AI systems during their whole life cycle. Not only the development, but also the everyday deployment of AI systems demand an accountable person - or group of persons - in charge for the processes. Regarding covert AI systems the draft states that “(AI) developers should therefore ensure that humans are made aware of – or able to request and validate the fact that – they interact with an AI identity” (p. 11). The relevance of this topic is exemplified by the ever-increasing use of chatbots (in either written or vocal communication) where it is not always obvious for the user that the communication partner is not a human. Identifying such non-human communication partners as such is a trust-building measure. Labelling chatbots as such and providing alternative formats of communication if exchange/contact with AI is not desired is an area worth looking at.	In general, we would like to suggest the consideration that the authors differentiate between B2B-based AI systems that are built for a professional context (i.e. flight controller, accounting, social security) and B2C-based AI systems for consumers (i.e. dating apps). The ethical framework and the requirements of Trustworthy AI (cf. p. 13) may differ significantly. Regarding the accountability of AI system the paper states that “good AI governance should include accountability mechanisms, which could be very diverse in choice depending on the goals. Mechanisms can range from monetary compensation (no-fault insurance) to fault finding, to reconciliation without monetary compensations. The choice of accountability mechanisms may also depend on the nature and weight of the activity as well as the level of autonomy at play” (p. 14). This section should also discuss how to handle accountability in the case of severe wrong decisions, e.g. such that cause the loss of human life (e.g. in Health, Autonomous Driving etc.). However, at this point, we would like to stress once again the importance of differentiating terms of liability from terms of responsibility more clearly. The section “Data Governance” (p. 14) describes the technical process and challenges of machine learning training, but not how to deal with data acquisition. Since AI is a data-driven model, the ultimate decision is who has the most and highest quality data. The higher the data quality, the better the AI algorithms. Bilateral agreements are necessary and a kind of data hub that functions as a control instance for the exchange of data be-tween nations. The paper also states that "When data is gathered from human behaviour, it may contain misjudgement, errors and mistakes. In large enough data sets, these will be diluted since correct actions usually overrun the errors, yet a trace of thereof remains in the data" (p. 14f). However, this might be too optimistic. One cannot rely on self-correction due to large enough data sets. Even a large set of data may contain a structural bias which can eventually be passed on to the products (bots, algorithms etc.) that are built from them. We would like to question the published statement that “systems should be designed in a way that allows all citizens to use the products or services” (p. 15). Yet AI products and fea-tures appeal to different target groups. A “One-design-fits-all”-approach does not seem to be very practical.	It should be emphasised that the steps presented in the section “Assessing Trustworthy AI“ can only be the first interventions on the road to create a Trustworthy AI. Their concretisation and implementation requires entrepreneurial initiative in order to implement them and to turn them into economic reality. We thus suggest to stress that organizational impact needs to be taken into account not only during design time, but also during runtime of AI systems. We would also like to point out that the list of questions currently appears to us to be provisional in some respects. The published questions should be - in order to refer to the introductory chapter of this document - seen as the starting point, not as the end of a discussion on the ethical aspects of AI.	Bitkom welcomes the two components of Trustworthy AI, stating that it should ensure an ethical purpose (e.g. to respect fundamental rights) but also be “technically robust and reliable since, even with good intentions, a lack of technological mastery can cause uninten-tional harm” (p. i; 6f). The published guidelines “are not meant to stifle AI innovation in Europe, but instead aim to use ethics as inspiration to develop a unique brand of AI, one that aims at protecting and benefiting both individuals and the common good. This allows Europe to position itself as a leader in cutting-edge, secure and ethical AI”. (p. ii) We strongly support the idea of protect-ing the wider public's interest with regard to upcoming trustworthy AI products and ser-vices. We thus suggest that the draft includes clear language so that the proposed guide-lines can be expected to achieve relevancy and protect the interests and rights of Europe-an citizens. The ambition should be to bring the guidelines fully into practice as/via do-main-specific ethics code(s). The focus of policy makers should be to strengthen the Europe-an AI ecosystem. In our opinion, the guidelines should also contribute to an understanding of the social learning process regarding AI and increase general trust in AI. As a new basic technology, AI creates a learning process for all stakeholders. In this process, different learning abili-ties, the willingness to learn and responsibilities have to be taken into account. This cer-tainly also involves consumers, which must also be given the opportunity to partake in this process. In the guidelines draft, however, the consumer mostly appears as an object to be protected (this is due to the "rights approach to be protected") and thus not as a self-empowered subject. This potentially neglects consumer’s role within the ethics of AI. We also want to state that it could be instructive to deal with the ethical aspects of AI on the basis of time scales. For example, the question should be considered whether AI sys-tems should assume full control and decision-making autonomy beyond human capabili-ties (e.g. under 500 milliseconds) below a certain time. What about hours or days when autonomous decisions can become reversible as more information becomes available? Just because there is a lot of time to reverse a decision, there may still be areas where we do not want the AI to make such decisions without human supervision. We would also like to draw attention to a general discussion which also concerns the

				In order to address the requirements to achieve Trustworthy AI, we would like to add another non-technical method (p. 18ff) to be employed within the development process: AI systems should come with a clear description of their limits, including the areas they are intended for and those they are not intended for, as well as a description of input data that the system cannot properly cope with (e.g. an animal recognition system that has been trained with data on mammals might not suit well for identifying insects). In our view, the distinction between "ethical purposes" and "technically robust and reliable" is necessary. However, the Draft does not clearly distinguish between those two aspects, especially from Chapter II onwards. However, this would make sense. The aspects "techni-cally robust and reliable", to which sections 8 (robustness) and 9 (security) are most likely to be assigned, should be expanded. Regarding "Figure 3: Realising Trustworthy AI throughout the entire life cycle of the system" (p. 18): Please put 'analysis' in the top left corner of the four boxes that indicate the recur-sive flow of actions to maintain and improve over the life cycles - it seems more natural in the mode of reading top to bottom, left to right, that analysis is the start.		meta-aspects of these guidelines: Within philosophical discourses on technology, the question is discussed to what extent algorithms can actually make decisions. The way we talk about AI has significant impact on this principle and should be taken into account by the guidelines. We propose to speak of "AI Processes" and not "AI Decision Making" or "AI Decisions" as these are key phrases within critics of AI. A more nuanced wording of AI concepts could possibly lead to a higher acceptance of AI by the public. We welcome the fact that the guidelines should not be seen as an end point, but rather as the beginning of a new debate on Trustworthy AI (p. 2). We furthermore like to emphasize the importance of national discussions (e.g. the Data Ethics Commission Ethic of the Ger-man Federal Government) on Trustworthy AI, which should also be facilitated and consid-ered by the paper. The paper also states that the final version of the document will set out a mechanism that enables all stakeholders to formally endorse and sign up to the guidelines "on a voluntary basis" (p. 2). This aspect raises important questions such as: If a stakeholder "formally endorses" the guidelines, would they substitute already initiated self-binding Codes of Conducts? If a stakeholder would not formally endorse the guidelines, would that create the impression that the stakeholder does not support ethical aspects of AI? And what role do national associations and their member companies play? On page 5, freedom occurs only in the limited form of "democratic freedom" and is quite insignificant overall. The idea of freedom aims at the responsibility of individuals and organizations. Orientation towards this value would therefore facilitate and promote individual and collective responsibility. The broad use of AI could help in this process, but poses also challenges in regards to avoiding responsibility and self-empowerment. This is a relevant challenge, which should be taken into account.
jesus	salgado	querytek	Rationale and Foresight The way it is expressed seems to the reader like ethics is the "tool" to reach the complete deployment of AI (the goal). I think it would be important to stress that ethical behaviour is the main goal. Trustworthiness is a good consequence. Like in the phrase "To ensure those benefits, our vision is to use ethics to inspire trustworthy development, deployment and use of AI. The aim is to foster a climate most favourable to AI's beneficial innovation and uptake." The goal is wellbeing of citizens, and wellbeing, good life, involves trustworthiness of, among many other things, the AI technology. In some parts of the chapter (and other parts in the document) it looks like "ethical AI" term is used like a kind of "branding" more than a true, honest approach to a safe, secure AI for humans. (looks like...I know it is not like that) Purpose and target audience Adresees are the stakeholders that develop, deploy or use AI. Does not seem to include other stakeholders that are impacted by it, although not using it or being customers,	In "testing and validating" and "Traceability & Auditability " I would include the need for CONTINUOS monitoring of unintended outcomes (like for instance a group discrimination) , since conditions in data and algorithms may vary (specially in reinforced learning algorithms). Much Like it is done in the first paragraph of chapter III. I think it is very important. In the Standardisation chapter, I would stress that, for each specific application, sector specific regulations may already apply. This should be included, since for healthcare, banking, aeronautics, market research and many other applications, a big corpus of specific international agreements should apply already. AI should first comply with those.	I would not know where to put it but I would include this sequence of questions: What is the goal/purpose of the AI application Who are the stakeholders affected (business, clients, government, public in general or groups of citizens) are all stakwholders and sub-groups of stakeholders aligned with the purpose If not what are the potential implications for these groups (from life threatening to just upsetting)	In the glossary: Human-centric AI: the phrase " the development and use of AI should not be seen as a means in itself, but with the goal of increasing citizen's well-being" should say better "...should not be seen as an end in itself, but …"

			like the citizens that are impacted by AI decisions. I think that citizens awareness about their rights and risks is a fundamental piece for the correct development of AI.				
Hervé	Falciani	TACTICAL WHITLEBLOWER ASSOCIATION	AI ETHICS MUST SERVE THE COMMISSION VISIÓN OR IT WILL LACK OF USEFULNESS The ethical framework of the IA should not be isolated from the other two pillars that underpin the Commission's vision: (i) increase public and private investment in AI to boost its acceptance, and (ii) prepare for socioeconomic changes, especially when according to some forecasts the AI can threaten around half of employment that must seek for other new occupations or livelihoods. Ethics must serve the fulfillment of these two objectives or it will lack of usefulness and human meaning or purpose.ACCURATE INFORMATION IS ESSENTIALIt will be essential to provide, in a clear and proactive way, accurate information to the parties. Faults or bias can limit the ability to make rational decisions, so they must be sanctioned and corrected from the ethical standards recommended to the society, as well as by public authorities.THRUSTWORTHLY AI INTELLIGENCE HAS INIFINITE USES As pointed out in the guidelines, trustworthy AI Artificial Intelligence not only helps improving our quality or more efficient delivery of healthcare services, promoting gender balance, tackling climate change, and helping us make better use of natural resources, but can also prevent fraud, tax evasion and money laundering, challenges for many financial organizations. Analysts estimate that AI will save the banking industry more than $1 trillion by 2030. (https://thefinancialbrand.com/72653/artificial-intelligence-trends-banking-industry/)AI HAS A GREAT POTENTIAL IN THE PREVENTING FRAUDAI has the potential to help the public administrations and financial systems become more efficient in the process of detecting tax evasions, fraud and money laundering. To quickly identify potential fraud, AI engineers have developed tools and systems that automatically aggregate and analyse data that normally requires many hours of labor in just a matter of milliseconds. For example Tactical Whistleblower Association, a non-for-profit association incorporated in Spain in 2018 that brings together experts in transparency, privacy protection, economics and social impact, antifraud, international commercial law, pure and applied mathematics, artificial intelligence, business, marketing, finances, Blockchain and many others has developed Taboow Project which operates as an autonomous rating and scoring platform acting as autonomous assistant. Its models for example, are fully auditable, which permit ethical control over automatically pondered criterias. For example, we all have seen major privacy challenges arising in the past few years can devalue the ethical standards, as a result Taboow has adopted a very specific approach: preserving users ´interest, along with preserving businesses interest. Using blockchain technologies privacy protection can secure operations while preserving confidentiality. Our AI algorithms are intended to add fake information to the public information when	INCENTIVES AND SANCTIONS WILL BE IMPORTANT IN COMPLYING WITH ETHICSThe promises of a better human well-being and the need to focus on the human being and respect for fundamental rights will not be achieved only with ethical recommendations of voluntary compliance. Sustainability is an objective expressed by the UN, now through its 2030 SDGs, and a principle enshrined in European treaties, for which it must bind its parties. The same can be said with European regulations and national and regional laws and norms. There is a lack of ethical guidelines on the legal compliance of AI.They must introduce incentives and sanctions aimed at enhancing the capacity of goodness of the human being as the supreme expression of his intelligence.There is a lack of a Sustainability Ethic in all its four main dimensions: intergenerational, environmental, social and improvement of the economic governess.	A NEW FRAMEWORK OF RESPONSIBILITIES SHOULD MAKE CLEAR PROVISION OF TRACEABILITY OF RESPONSIBILITY AND SAFETY AT EVERY STAGEConcerning “Traceability & Auditability”: In a view to ensuring a fair balance between the interests of producers and of consumers, a new framework of responsibilities should make clear provision for the traceability of responsibility and safety at every stage of the product value chain and throughout its estimated lifecycle, incorporating sustainability as a new factor that will make product updating, improvement, portability, compatibility, reuse, repair or adjustment a requirement. (EESC own-initiative opinion on IoT).	EU PRINCIPLES SHOULD BE INCLUDED IN AI ETHICS To the principles mentioned in point 4, it is necessary to add others incorporated into European treaties, national constitutions and international agreements, especially those that can guarantee access or inclusion and avoid exclusion in their various firms: in addition to gender, intergenerational, Digital, financial, race, ideology, etc ..,Trustworthy AI should implement reciprocal analysis and automated ethical assessment models to dynamically and directly prevent itself from employing unethical criterias.Generation of auditable AI models is a first prerequisite for an AI to be trustworthy.A second prerequisite is for the AI itself to be able to evaluate its own models.Auto-correcting principles are required and one of the fundamentals of Taboow´s engineered AI.	ETHICS SHOULD GUIDE THE DISTRIBUTION OF ADDED VALUE GENERATED BY AI TECHNOLOGIESThe same rules should apply to everyone to access to information, data, knowledge, markets and public authorities.Ethics should guide the distribution of added value generated by technologies.Equality remains the cornerstone of the #2030 Agenda, ensuring the rights & dignity of everyone, no matter where they are from, what they look like, who they love or what their socio-economic status is.EMERGING DECENTRALIZED TECHNOLOGIES LIKE BLOCKCHAIN COMBINED WITH ETHICAL AI CAN SOLVE SECURITY AND TRUST ISSUESTaking into account the concentrations of power and information in private entities that collect and process personal data, the increase of transparency on data management as well as the adequate protection against any infringement of the fundamental rights of citizens is needed. Emerging decentralized technologies like blockchain combined with ethical AI can solve security and trust issues: this can be used to track sensor data measurements and prevent not only duplication with any other malicious data but also safeguard the integrity and traceability of changes. (EESC own-initiative opinion on IoT).Decision making process requires more and more justifications to support it. As machine learning and intelligent assistants are increasingly inserted in the decisional process, the less we usually know of how forecasting and decision support were provided. With Prescriptive Artificial Intelligence, this would be no longer the case and assistants’ decisions audit and assessment is needed. The algorithms should be coupled with data provenance tools that offer you the ability to investigate and trace the decision process to ensure versatility, adaptability and audibility for all the technologies it provides.

		re-identification algorithms detect a risk of involuntary disclosure of information. IA IS AMONGST THE NEXT GENERATION DISRUPTING TECHNOLOGY FOR CONSUMERSMoreover as mentioned in the guidelines, AI—including the Internet of Things (IoT) and image recognition—is among the next-generation technologies that are considered to be disrupting for business and public governance models but also for consumers.SMART CONTRACTS WILL PLAY AN IMPORTANT ROLE IN AI ETHICSIn addition of smart contracts, which serve as blockchain's governing laws, can be also be an automated way to ensure ethical rules agreed upon between AI developer and individual data contributor are enforced.
MELISSA	COUTINHO	HLEG on Artificial IntelligenceThe Draft Ethics Guidelines for Trustworthy A.I. is an excellent first draft, with some obvious omissions highlighted by way of questions. That it invites comment and feedback is commendable. However, it makes a number of assumptions also, e.g. that Ethical Guidelines are a good starting place, and implies that the concept of "ethical" may be something that A.I. can be 'taught'. I have some qualms about that which I will detail in this response drawn from my own experience of working with innovative A.I. and its myriad, (mainly medical in my field) uses. Artificial intelligence & AccountabilityAccountability is touched upon in the guidelines as a principle but not focused upon. This may be a matter of language and terminology rather than non-recognition of the importance of this concept. The fact remains is that it may be worth elaborating upon. Expansion would be useful in terms of where accountability is required, e.g. is it the manufacturer, or user, or combination thereof, or does it depend on the type of issue involved? There are obvious distinctions that could be usefully made.New technologies and inventions have regularly given rise to issues regarding control, historically, looking at the motor car and dynamite, and more recently, explosive substances, and drones. Misuse and abuse may seem like similar words but can be used to describe different concepts depending on whether wrongful use is deliberate or accidental. In both instances, liability may need to be distinguished. Counter creations have been borne out of necessity with tracker and radar technology coming into their own, once arms carrying aircraft became commonplace during war.Addressing misuse/abuse is a necessary pre-cursor to being able to consider an item trustworthy. The motor car is a good example when first commercially available, of the issues concerning poor manufacture, maintenance and use. This could be a useful model for future A.I. developments. This could mean that failures linked to the three stages of the product life-cycle are treated distinctly, even if the product has virtual as well as physical

components. In this way, defects in manufacturing are for the manufacturer or equivalent. Failurers for maintenance and use are down to the user, with insurance held by both parties to cover their separate liability. The difference would be their education, or rate of learning, and their may need to be safeguards in terms of how broad connectivity is, in terms of machines communicating with each other so that a single error does not create widespread problems and that there is more sophistication in learning, where subjective decisions and opinions are considered information. Opinions can be considered important data, such as voting scandals and data misuse indicate. While the paper mentions that many products success depend upon a healthy respect and consciousness of what might happen if they are misused, this cannot be an over-riding consideration. To operate in this way would mean that commonplace items like knives would not be used, which are essential to everyday life, even if their use by many in violent scenarios is deplorable. Or cars, simply because one could be used as a weapon in a deliberate attempt to cause harm. For A.I. the scale is an issue, given that several thousand views worldwide, might influence A.I. in a negative way, (e.g. the virtual teenager invented by Microsoft who needed to be killed when her "learning" proved defective: (https://www.telegraph.co.uk/technology/2016/03/24/microsofts-teen-girl-ai-turns-into-a-hitler-loving-sex-robot-wit/). In such a scenario, volume was more important than other criteria, with bias from some proving a great influencer. The more power such a piece of A.I. possesses, the more the capability of harm on a large scale.Scoring/Quantification is highlighted in this document, but this does already occur, along with categorisation for social/health care by governments. Social security payments, particular healthcare entitlement, fertility treatments, social care assistance is the norm for many people in the EU already. (Some of these are standard agreements, other social constructs, and some allowed for in law, such as power of attorney for welfare – will this include A.I. constructs, when such agreements were entered into, before the existence of such sophisticated technology?)Checking that data recorded can be utilised properly – what counts as successful outcome, (e.g. if it is a young person surviving with all faculties intact, then will this be prejudicial for an older person without all faculties operating?) So the old premise of using the correct language for technology and its applications is important. A real case of what in computing circles is known as Garbage In; Garbage Out, with poor data, meaning outcomes are also poor.Ethical overrideA variation of the "Stop button" might involve a simple override. Who has access to this, and whether it is a regulator needs to be assessed. Access is allowed in comparable scenarios, such as cosmetics formulations to Poison Centres to address accidental ingestion or urgent treatment notwithstanding the commercially sensitive intellectual property involved. If appropriate Regulating Agencies exist, should they have overrides, or should this be given to specialist sectors of the Emergency Services? (It would certainly have been

helpful in countries where drones at airports meant that hundreds of thousands of people missed flights last year!)Teaching of A.I.Teaching A.I. ethics: who does this and what responsibility for this will be set out? Will A.I. learn from available information that exists on the world wide web, which can include extremist views/positions. Will the philosophical and ethical positions that we understand to be correct be accepted by A.I. or could they draw different conclusions, on what eliminating the poorest or least successful can achieve in objective terms?Annual MOT/ Assessment This could include assessment of override and learning componentsLike cars and their annual checks, or other products which require Notified Body oversight, could there be certification, which is limited until regular checks are carried out. Could these be done by any approved Regulator or person with expertise in identifying and fixing relevant issues, such as the equivalent of mechanics and car repair garages?While such expertise would itself need to be assessed, there should be sufficient safeguard, particularly in the early years to ensure that maintenance and assessment takes places alongside innovation. Assessment need not mean stifling creativity, as the car and aircraft industry illustrate, albeit that some models may be found to be not suitable commercially, or require restricted use.Limiting communication between machines might be a component, so that some forms of machine learning/data access can be switched off? Distinguishing fact from opinion needs to be considered also? Consider what extrapolation might give rise to by way of conclusion, e.g. would data about certain races being more successful than others historically in certain areas, lead to increased discrimination in companies where there has been historic inequality but commercial success, leading to problems being repeated rather than resolved?A limiting concept of good/bad; if this means that there is a to be a risk/benefit assessment then it may consider that there are overall good outcomes than nonetheless allow harm? This will be the case in many medical situations, given the vulnerability of a person with a disease/problem, as little is risk free. Further, new cancer treatments and experimental treatments, particularly those linked to genetic may not be able to guarantee overriding benefit compared to rusk; they will still be the only option for some patients who are prepared to take the risk of harm given their limited alternatives, ignoring the absence of data to confirm a positive benefit:risk profile. Further, the concept of risk:benefit is not always correct on a global or sector basis but may be accepted in an individual case.In terms of an assessment list for Healthcare Diagnostics and Treatment, these can already be derived in part from Good Medical Practice, and Good Manufacturing Practice. There are already Ethical Boards that allow for assessment of new products to market and their checklist could simply be varied. Further the electronic component of A.I. is already contained in multiple EU documents along with the necessary standards for producing and protecting good data; elements of all of these would additionally be needed for an appropriate and comprehensive check-list. Lastly, the concepts are as important as the

technicalities, as the unknown cannot be completely addressed but can be prepared for, with advance planning and some conceptual thinking.ConclusionHappy to look at specific issues or problems that are flagged up for me and to address these further. Commend to the group the Big Data work that has been begun/done at EU level. Whether it is sufficient for these ethical principles to be simply Guidance or this needs a different classification to reflect the importance of signatories: e.g. a Convention, worth considering. While this is not EU centric, A.I. not geographically bound in the same way that other products are and putting the genie back into the bottle is a task that is unlikely to reside with one part of the world alone but be a global endeavour.Melissa Coutino(Medical Devices Lead Lawyer to MHRA for 10 years, Principle Negotiator for UK for Medicrime Convention.)

Hélène	Benoist	European Banking Federation	Introduction: Rationale and Foresight of the Guidelines The EBF welcomes and supports the acknowledgment in the introductory sections of the Guidelines that this is an emerging area and that changes to the Guidelines will be necessary over time. What is deemed "ethical" varies between individuals, societies, and jurisdictions, and can change over time. It is important to recognize that the purpose of ethics is to help decide what is right or wrong, which is best accomplished through a set of abstract, non-binding guidelines. Ethics bridges the gap between the regulated and non-regulated spaces — that is, firms know what they can do in accordance with relevant laws and regulations, but ethics guides firms on what they should do. Firms need to be able to come to the decision themselves about what is right or wrong, beyond legal and regulatory requirements. The guidelines should provide firms with practical information and tools, without overlapping with existing requirements, to consider all factors relevant to the activity and outcomes, and to draw an informed, ethical conclusion. As such, we appreciate that the start of the guidelines recognises that different contexts will require different approaches, with flexibility required in application (page 3, "Scope of the Guidelines"). While acknowledging that different contexts require different approaches, it seems crucial to ensure ethical standards are consistently applied across technologies.We believe this principle could be better reflected in the body of the Guidelines which takes at present some strong positions (e.g. there are numerous statements that certain things must or must not occur). Given the preliminary and evolving nature of this guidance, it would be more appropriate to soften these points to encourage relevant stakeholders developing, deploying or using AI to turn their minds to important challenges and ethical principles, with some flexibility in how to approach these, instead of suggesting hard rules.We also support the higher intent of the Guidelines to foster reflection and discussion on an ethical framework for AI at a global level. To achieve this goal, the Guidelines should fully consider the subjective nature of ethics and differences across cultures in order to have a framework that can provide a suitable basis for broader discussion.We would suggest emphasising the need for flexibility in application throughout the guidelines and have thus made some suggestion for amendments below.Furthermore, ethical standards need to be technology agnostic and should not set different standards for different solutions. Specific ethics guidelines for trusted AI have the risk that AI-based processes are subject to higher standards than conventional human-based activities and are thus discouraging the uptake of AI. In any ethics guidelines, a definition of AI must also be considered very carefully as it would draw the line between an organisations' processes where ethics guidelines are encouraged to be applied and those other areas which are out of scope by the definition of AI. As a general matter, we suggest that any ethics standards should apply to all technologies and not set different standards for different solutions. This is critical also because there is no commonly agreed definition of AI. For example, the definitions rely on	As an overarching comment, we suggest reframing this section. Compliance with regulation and respect for rights do not, in themselves, give rise to a "purpose". We suggest rewording the heading to focus on "Ethical Intent". The guidelines define Ethical Purpose as AI which ensures compliance with fundamental rights and applicable regulation. The topic of ethics should be approached as guidelines over and above regulation and rights, which differ by country. We thus propose the following amendment:"Respecting Fundamental Rights, Principles and Values – Ethical Intent"However, a clear statement of "Purpose" for every AI system could bring useful benefits. See also comments below on Chapter II.On the part on "Ethical principles in the context of AI and correlating values", although we agree with the principles presented, we would suggest some amendments in order to ensure some level of flexibility. Please see these in more details below. • The Principle of Beneficence: "Do Good" (page 8): We note that many positive use cases of AI will be for commercial purposes. While the draft Guidelines acknowledge the potential benefits to the cost and quality of services, this point should be strengthened to make clearer that AI can make a positive contribution through commercial innovations.Furthermore, we note that "wellbeing of the user" as defined by the user may sometimes contradict with other objectives, rights, values and principles. Although it can be a part of the input provided into a system's functionality, the system cannot (and probably should not) ensure that user-defined wellbeing is fully met. As outlined in more detail under "Do no Harm", the Guidelines should recognise that a careful balance should be struck between the benefits derived from an AI use-case and the potential harms.• The Principle of Non maleficence: "Do no Harm" (page 9): Although we agree that technology should not be created with harmful intent, we would suggest amending this part slightly. It is appropriate for "harm" to be defined broadly so that firms consider the risks of a potential new use for AI carefully. However, given the very general definition of "harm" presented in this part of the document, saying that no individual will ever be "harmed" and that in no circumstances could this be justified will mean that AI – or at least "Trustworthy AI" adhering to the ethics guidelines – will be severely restricted and use- cases of potential benefit to other individuals or to society will be prevented. Furthermore, it seems highly unlikely that there will ever be a situation where all harm is completely avoided and it should be noted in the guidelines that harm to certain people could be interpreted as benefits to others. As such, this principle would require some tolerance and boundaries in order to be able to realistically operate, especially given that the notion of "AI specific harm" is not defined. In practice, efforts to avoid harm need to be balanced with other goals, such as achieving justice, or helping people ("beneficence"). Instead of seeking to "prevent" any harm, firms should carefully identify and consider potential harms, and benefits. We believe the principle of "do no harm" should be for potential harms to be carefully balanced against the positive benefits of the technology. The firm should carefully balance these against each other to	Chapters II and III go beyond principles and appear like a high-level compliance document. Detailed rules, even if only recommended best practice, need to be carefully considered in order to prevent unintended consequences. Given the diversity of AI use-cases, a 'one size fits all' approach will not always work. There may be multiple means of achieving a similar outcome and it should be clear how to proceed in case a given recommendation cannot be met or is not applicable in a certain case. We therefore suggest:§- Reframing the ten "requirements" as "key considerations" to clarify that this is guidance to inform firms' approach to ethical AI development, rather than a set of compliance obligations; §- Clarifying that these will need to be adapted by firms to their specific use-cases.This would be consistent with the approach in Chapter III on Assessing Trustworthy AI, which states "Moreover, the precise questions will vary from use case to use case, and a tailored approach needs to be taken for each specific situation, given the context-specificity of AI".Furthermore, we suggest that, considering the importance of this part on Trustworthy AI and the following chapter on Assessing Trustworthy AI, and taking into account the fact that these guidelines aim at being a living document which evolves with the technology and its understanding, we would suggest this part of the paper be adopted in its final form at the same time as the Policy and Investment Recommendations. Given the tight timeframes for completion of the Guidelines, it would be sensible to conduct an additional consultation on Chapters II and III to ensure that the diverse impacted sectors and interest groups have an opportunity to provide input. This consultation could be launched at the same time as the finalisation of Chapter I. At least, we would welcome an introductory paragraph to both Chapter II and Chapter III restating that these Guidelines are intended as a living document, bound to evolve with an increased understanding and knowledge of the technology.In line to what has been commented on the "do not harm principle", the requirement to "not use data against the individuals who provided it" should be clarified as there may be legitimate actions which could be considered as being "against" the individual who provided the data while still in line with regulatory requirements and socially beneficial, etc. (e.g. fight against terrorism financing).Please see also below some other parts we believe could be further extended or considered as a part of ongoing work on Chapters II and III.§- Data Governance (page 14): We would welcome a suggestion in this section that having good data quality and representative data sets are important considerations as both data quality and representative data sets are prerequisite to minimize unfair bias and discrimination.§- Design for all (page 15): As currently drafted, the paragraph may be complicated to implement in practice. Further work and clarifications on the "design for all" approach would be welcome.§- Governance of AI Autonomy (Human Oversight) (page 15):Questions over accountability appear frequently in the Assessment chapter, which indicates that clarity over accountability is seen as an important test for Trustworthy AI. However,	Process:Considering the importance of this part on Assessing Trustworthy AI, as well as the previous chapter on Trustworthy AI, and taking into account the fact that these guidelines aim at being a living document which evolves with the technology and its understanding, we would suggest this part of the paper be adopted in its final form at the same time as the Policy and Investment Recommendations. Chapters II and III go beyond principles and appear like a high-level compliance document. Detailed rules, even if only recommended best practice, need to be carefully considered in order to prevent unintended consequences. They should also be clear how to interpret them and how to proceed in case a given recommendation cannot be met. Given the tight timeframes for completion of the Guidelines, it would be sensible to conduct an additional consultation on Chapters II and III to ensure that the diverse impacted sectors and interest groups have an opportunity to provide input. This consultation could be launched at the same time as the finalisation of Chapter I.At least, we would welcome an introductory paragraph to both Chapter II and Chapter III restating that these Guidelines are intended as a living document, bound to evolve with an increased understanding and knowledge of the technology.Notably, the parts on "Respect for Privacy" would benefit from greater clarity, with care taken to align with GDPR and avoid duplication. We would suggest including detail around controls and having a clear basis for processing the data. There are only a few brief mentions of data governance and appropriate controls through the draft guidelines (i.e. page 22 "Accountability Governance" and "Codes of Conduct", as well as their respective assessment list in Chapter III, page 25 and 26). Care should also be taken to not broaden the implied scope of the GDPR (e.g. suggesting that all derived or inferred data is personal data, which is not the case). Further clarity, in collaboration with the policy and investment recommendations, on the issue of data governance would be welcome.Additionally, this chapter contains "questions that should be reflected on", but it is not clear how these are to be applied in practice. The guidelines suggest that a tailored assessment list will be produced for four particular use cases, and it will be useful to see these lists once they have been prepared.In addition to this "customisation" to the appropriate sector, the guidelines should make clear that these will need to be adapted by firms to their specific use-cases and that the specific features of each AI solution will also determine which questions apply and what approach is proportionate, such as the impact of decisions made by the AI; scale; nature of data being processed; third parties involved and so on.	Interactions and overlaps with the GDPR:As stated above, the guidelines should be seen as guidance to inform firms' approach to ethical AI development, rather than a set of compliance obligations. As such, and with regards to data protection questions, we believe it useful to refer to the work the European Data Protection Supervisor (EDPS) has conducted on the issue of digital ethics . Indeed, the 2018 report of the EDPS Ethical Advisory Group (EAG) states the following: "The EAG expressly avoids an instrumental approach to ethics of a kind that would result in an ethical checklist or set of measures that, once accomplished, would essentially exhaust ethical reflection and release its practitioners from further discussion. The EAG wishes to discourage approaches to ethics governance that equate data protection with the application of do's and don'ts. On the contrary, it seeks to encourage proactive reflection about the future of human values, rights and liberties, including the right to data protection, in an environment where technological innovation will always challenge fundamental concepts and adaptive capabilities of the law. It seeks to inspire all relevant stakeholders to identify the areas where ethical problems not only emerge from the development and operation of today's digital technologies, but integrate in both their designs and business planning reflection about the impact that new technologies will have on society, generating their own guidelines for addressing them tomorrow while remaining vigilant to what their own guidelines had not foreseen, when by all accounts the premise, aims and impact will be astonishingly different from today." The EBF believes this approach to be appropriate, for the reasons mentioned in the EAG's report, and thus encourages the AI HLEG to refer to it. We have nonetheless identified a few key areas of overlap between the guidelines and the General Data Protection Regulation (GDPR), on which you will find more details below. Interactions and overlaps with GDPR – consent:As noted above, we are of the view that the ethics guidelines are not the most appropriate place to provide clarifying guidance on legal and regulatory requirements, such as GDPR. However, we have noticed some considerable areas of overlap between the Guidelines and principles and requirements of the GDPR. It would be useful to clarify these interactions. In particular, the approach to "consent" and "basis for processing" (GDPR Article 6) should be clarified.The Guidelines refer in several places to "consent", notably at the bottom of page 5 and top of page 26 in the context of GDPR compliance. It is unclear whether "consent" here is intended to refer to consent as a "basis for processing" under GDPR (as described below) or whether it is intended in a more general sense that individuals should not be compelled to buy a product, subscribe for a service etc. GDPR-style consent has a specific meaning:"Consent" is not required for valid processing of personal data under the GDPR (it is one of the legal bases of processing provided) and has a very particular meaning. Briefly, GDPR standard consent is only valid if the individual has a genuine option to refuse to consent without losing access to the main service, with the option of withdrawing the consent at will without losing access to the main service. If an

anthropocentric concepts of “perceiving”, “interpreting”, “reasoning” and “knowledge” which require separate unpacking before the definition can be operationally useful, i.e., capable of distinguishing AI from non-AI based processes and systems. As it stands, the border between what is considered non-AI and AI based process and system remains blurry, and although the High-Level Expert Group has proposed a definition of AI, this is not universally accepted and is likely to evolve. Regulating AI systems might be similar to regulating human behaviour, since they have some sort of autonomy, and adapt and learn due to the nature of machine learning. Therefore, it is very difficult to pre-certify machines at design time as being ethical or guarantee that they do not misbehave when applied in real world. Similar to human judgement of ethical behaviour, we should set expectations, observe the behaviour and hold someone accountable for misbehaviour. Finally, we understand the guidelines aim at being quite high-level, rather than setting out detailed rules. This is appropriate, but one consequence is that it could be hard to identify in advance how they will be applied to real scenarios. More generally, given the broad scope of AI applications, it is possible that the guidelines could inadvertently constrain some beneficial use cases; for instance:§- The beneficence and non-maleficence sections could seem to restrict a wide range of potential uses, as could the discussion on bias and fairness;§- The sections on transparency and explicability should reflect the fact that some AI applications will require higher standards of transparency than others, particularly for use cases aimed at detecting fraud, money-laundering, etc.;§- The references to consent do not allow for circumstances where it is not appropriate to rely on consent, or where no personal data is involved;§- The guidelines do not reflect the fact that program code and techniques can be valuable commercial intellectual property, requiring protection (and not be open to “inspection” from third parties).We expand on these points in the more detailed discussion below where we provided some recommendations and suggested some amendments.

determine whether the potential harms are justified, given the benefits anticipated from the AI system. This would also be more consistent with the principle of Justice. For instance, an agency’s trading algorithm that allows a client to execute a trade with minimal market impact will benefit this client, at the expense of other market participants who may be seeking to detect that client’s trade and trade against him or her. This would also align to the principle to “be fair” which talks about balancing positives and negatives for different groups. There may also be situations where an AI system needs to refuse AI services to an individual, for example, thanks to controls that ensure suitable or responsible use of the AI system.We would thus suggest the following amendment: "At the very least, potential benefits of AI systems (and similar technologies) should be considered keeping in mind the potential harms of the technology".AI HLEG draft ethics guidelines for Trustworthy AI, Page 9• The Principle of Autonomy: “Preserve Human Agency” (page 9):The document presents human “autonomy” as always desirable, with the requirement for a human override. We note that there will be cases where a human broad override does not make sense: for example, in safety systems, or, in financial services, in the detection and prevention of bad conduct, or the prevention of dangerous trading decisions. In these situations, there should be human oversight and accountability for the AI system, and individuals might need the authority to request the review of an automated decision when they have compelling cause. However, the individuals interacting with the AI system should not be able to directly override the AI system without cause, as this would undermine its protective function.• The Principle of Justice: “Be Fair” (page 10): Bias has unfortunately been prevalent in all societies and systems, well before the advent of AI, and no means have been found that would ensure that all individuals remain free from all kinds of bias or that the positives and negatives resulting from AI are evenly distributed. As a consequence, historic data sets used for training of AI systems also include biases . Based on this assessment, we believe that the appropriate test would be whether the AI leads to less unfair bias than an alternative system would, as expecting absolutely zero bias will not only not be operable but will prove harmful to the development of AI. AI technology should be seen as a chance to reduce unfair bias in future.Similarly, the definition of bias makes the assumption that a predominant way to inject bias can be in the collection and selection of training data, which is not necessarily true. Indeed, the selection of data-processing techniques is as important. Furthermore, intentional bias is sometimes included in algorithms for legitimate purposes and preventing this would limit the justifiable development of these algorithms. Indeed, sometimes we want “legitimate bias” (e.g. diversity in hiring). We suggest the wording be softened to be use-case driven, with documented controls on unintended bias. This would also be in line with some of the draft Guidelines recommendation (e.g. to ensure the teams developing, implementing and testing AI products and solution are diverse and inclusive).Instead, we recommend the Guidelines to refer “unfair

accountability is less clear in the previous sections, where the emphasis is on oversight.On page 14, “Accountability” is described only in terms of redress for the wronged party. Firms who develop and deploy AI will most likely need to experiment with different governance approaches which would work for them based on their size, organisational structure, the type of AI applications, etc. Governance approaches could take several forms (e.g. an – or several – “accountable person(s)” for specific AI projects), but these should be up to the individual firms to figure out. §- Respect for Privacy (page 7): This section would benefit from greater clarity, with care taken to align with the GDPR and avoid duplication. We would suggest detail around controls and having a clear basis for processing the data. There are only a few brief mentions of data governance and appropriate controls through the draft guidelines (i.e. page 22 “Accountability Governance” and “Codes of Conduct”, as well as their respective assessment list in Chapter III, page 25 and 26). We recommend to also take care to not broaden the implied scope of the GDPR (e.g. suggesting that all derived or inferred data is personal data, which is not always the case). Further clarity, in collaboration with the policy and investment recommendations, on the issue of data governance would be welcome. §- Statement of purpose:Compliance with law, regulation and fundamental rights does not, in itself, constitute an “Ethical Purpose”. Similar to the approach under the GDPR, an option for firms to consider as a part of their AI governance could be to ensure that AI systems have a clear statement of “purpose” that the system is trying to achieve. This could be accompanied by a description of measures which the AI designers have sought to optimise in order to achieve that Purpose.This Purpose would make clear to users / subjects of the AI system what it is trying to achieve. It could also be a tool for the firm to document its intentions vis-à-vis auditors or regulators. The requirements for accountability and auditing should be tailored based on a classification of the AI models with respect to their potential impact and risks (e.g. requirements for marketing models can be different than those for risk estimation models). §- Diversity in setting up teams developing, implementing and testing the product: The EBF welcomes and supports this recommendation of ensuring the teams developing, implementing and testing AI products and solution are diverse and inclusive. Indeed, this concept is extremely important as it is a key way of ensuring alignment with the “five principles and correlated values” detailed within the paper. Although it may not be a panacea, the concept is extremely important when thinking about bias and the need for “injected representative bias” which is required if we are to have a truly human-centric approach to AI. It’s no good injecting bias if it only represents the bias of a particular cohort of individuals.

individual must “consent” as a condition of gaining access to a service, the consent is not valid. As such, for processing that is necessary for a service to be provided, it is not valid to seek consent. Instead of relying on consent, under GDPR firms can also legitimately process personal data in five other circumstances under Article 6 (the firm must have a valid “basis for processing”). In financial services, these are primarily:§- Where the processing is necessary to enter into or perform a contract§- Where the processing is necessary to meet a legal obligation §- Where the firm has a “legitimate interest” in the processing (provided that any negative impacts on the individual do not outweigh the benefits of the processing)Under GDPR, consent is not an appropriate legal basis for processing many services:In the context of financial services, therefore, consent would not generally be sought for personal data processing that is necessary to provide a bank account (such as processing transactions) or to comply with regulation (such as monitoring for money laundering). Such processing cannot be freely “switched off” by the customer.If the Guidelines force firms to rely on consent as their GDPR basis for processing for AI processes, this would in effect make it impossible to develop a service or product that relies on AI in order to function, and especially complex for compliance matters. This could severely inhibit, for example, the use of AI to detect and prevent fraud and money laundering, given that firms would not reasonably be able to turn this function off on customer request without leaving fraudsters with the option to operate on a less efficient system; thus effectively leading to significant social harm.Where consent from the data subject is not sought, GDPR nonetheless requires a suitable description of the data collected, how it is processed, parties it is shared with, etc. The individual can then choose whether or not to buy the product, request the service, etc., but this is not “consent” for GDPR purposes.Furthermore, GDPR Article 22 limits firms’ use of automated decision making and grants individuals enhanced rights in relation to automated decisions. In particular, there is a right for individuals to have such decisions reviewed by a human being in many circumstances. This is likely to be highly relevant to AI systems.Providing clarity in the Guidance: In order to avoid confusion, the references to consent in the Guidelines should be clarified, with the focus changed instead towards being transparent about AI use to enable individuals / users to decide freely whether or not to use a service that uses AI. Similarly, it is unclear why consent is specifically referred to on page 26, given that this is only one of six “bases for processing” permitted under GDPR (and, as mentioned above, this is only applicable where personal data is being processed). This reference should be deleted or replaced with a reference to having a valid “basis for processing” under GDPR and meeting other obligations such as GDPR transparency requirements, and perhaps the requirements of Article 22.Interactions and overlaps with GDPR – right to be forgotten:In the data governance section, it is stated that “It is advisable to always keep record of the data that is fed to the AI systems”. However, this principle could be against the right to be forgotten recognized in Article 17 of GDPR,

bias", such as racial prejudice, as distinct from "legitimate bias" such as aiming to achieve greater diversity in hiring new staff. We would thus suggest the following amendments:"For the purposes of these Guidelines, the principle of justice imparts that the development, use, and regulation of AI systems and alternative technologies are fair. Developers and implementers need to ensure that individuals, especially under-represented communities, maintain freedom from unfair bias, unfair discrimination and from stigmatisation. AI should lead to less unfair bias than an alternative system. Additionally, AI systems and alternative technologies should avoid placing vulnerable demographics in a position of greater vulnerability and strive for equal opportunity in terms of access to education, goods, services and technology amongst human beings, without unfair discrimination. Justice also means that AI systems and alternative technologies must provide users with effective redress if harm occurs, or effective remedy if data practices are no longer aligned with human beings' individual or collective preferences. Lastly, the principle of justice also encourages those developing or implementing AI and alternative technologies to be held to high standards of accountability. Humans might benefit from procedures enabling the benchmarking of AI performance with (ethical) expectations."AI HLEG draft ethics guidelines for Trustworthy AI, Page 10We also note that an AI system might, despite best efforts, contain some undesirable bias but still less than the bias observed when the same decisions are made by human beings. In this scenario, we would argue that this could be a positive use of AI despite the existence of bias. It should also be noted that it might sometimes be difficult, if not impossible, to detect unintended biases or discrimination. Detecting unintended biases would for example require to collect sensitive attributes protected by the law (such as age, gender, race or religion) in order to ensure that no correlations to other attributes incorporated in the model exist. Requesting such data would raise additional ethical and data protection questions. We highlight that, while minimising unfair bias is important, it also poses a practical challenge: as soon as the discriminating attribute is correlated with the target variable all other attributes that are useful for predicting the target variable will also be correlated with the discriminating attribute. Hence, it is statically impossible to avoid discrimination a priori. Nonetheless, averaging over discriminating variables a posteriori may remove unintended biases . However, this requires the discriminating attributes to be defined and cannot be done if the discriminating attributes are unknown (e.g. sexual orientation), which requires collecting this additional, potentially legal protected, information . It would thus not be possible for a firm to demonstrate that it had prevented all possibility of bias in its AI system. Rather than expecting firms to prevent all incidents of bias, firms should have procedures in place to identify potential bias in AI systems, consider the fairness implications and take appropriate steps to ensure that overall fairness is achieved. This could include identifying clearly unacceptable / unfair bias, and also other types of bias that require further review. We also note that data quality is important in ensuring and the obligation to limit the time personal data is stored under GDPR Article 5.Vulnerable demographics: We agree with the broad statements made in the guidelines that the use of AI should have a positive effect on inclusion and diversity.This is a complex area that would benefit from further consideration, in collaboration with a broad group of stakeholders, after the completion of these guidelines. On the Definition of AI: We would welcome further clarity on the boundary between AI and non-AI algorithms, (e.g. with detail on "reasoning" in definition of AI). The definition document contains very useful explanations of different types of AI technology. However, additional detail should be included on the key concept of "reasoning", which is part of the updated definition of AI. In addition, examples of technology that would and would not be captured by the proposed definition would be helpful.On page 2 of the document, and in the final definition on page 7, the text refers to an AI system "reasoning on" the knowledge it derives from its environment. We would interpret this as meaning it carries out its "own" reasoning based on a model that it has built (i.e. by examining all existing data and deriving rules). However, it could be argued that an algorithm that applied a manually programmed set of conditions to such circumstances would be a form of "reasoning" – even though most would not consider this to be "AI".It would therefore be useful to clarify this in the definition (i.e. that an algorithm applying pre-programmed criteria / conditions does not constitute AI), as well as including further illustrative examples.Process:To ensure that the Guidelines provide a realistic and applicable framework ensuring the development of Trustworthy AI in Europe to all relevant stakeholders that develop, deploy or use AI, the EBF would support the submission of the final guidelines to a wider and longer consultation process. In particular, Chapters II and III go beyond principles and appear like a high-level compliance document. Detailed rules, even if only recommended best practice, need to be carefully considered in order to prevent unintended consequences as well as clarification on how to interpret them and how to proceed in case a given recommendation cannot be met. Given the tight timeframes for completion of the Guidelines, we feel it would be a sensible step to conduct an additional consultation on Chapters II and III to ensure that the diverse impacted sectors and interest groups have an opportunity to provide input. This consultation could be launched at the same time as the finalisation of Chapter I.It would also be useful for the Commission to produce a "marked up" version of the revised Guidelines in due course to help firms and the public identify changes.

fairness. Another important factor to drive out biases is diverse and inclusive teams building and testing the algorithms as this will help to identify additional unfair biases in historical data.• The Principle of Explicability: "Operate transparently" (page 10): We agree that transparency is key to building and maintaining citizens' trust in AI systems. However, the document presents transparency as always desirable. We believe transparency should not be so detailed as to undermine the use of AI in certain circumstances. It is indeed crucial to find the right degree of transparency vis-à-vis individuals, competent authorities, jurisprudence, etc. Transparency goes hand in hand with a loss of intellectual property and must therefore be well balanced. We would suggest to add wording providing exceptions to the principle of business model transparency as there exist some situations where opacity is needed, looked for or even has no impact (e.g. using AI for adapting user interfaces to customer preferences). For instance, in the financial services sector, AI is used to prevent and detect fraud, financial crimes or terrorism financing or cyber security incidents. Exposing how the technology works could allow the system to be gamed and risk undermining the (socially beneficial) purpose of the AI system. Additionally, a balance needs to be struck between this part of the document and intellectual property rights and trade secrets. The "baseline parameters" referred to in the paragraphs (which is only vaguely stated) would need to be defined carefully as requesting evidence of baseline parameters in addition to inputs of AI system would require exposing parts of, or entire, business utilities. This would need to be studied more carefully as it could hurt investments in AI and the development of the technology in Europe. Furthermore, the Principle of Explicability states that "Explicability is a precondition for achieving informed consent from individuals", which appears to combine the two concepts, as well as suggesting that consent should be the only basis for using AI (which is too restrictive): it would be preferable if the guidelines a) clarified references to "informed consent" to allow for AI use-cases not involving personal data or use-cases having no potential negative impact on the user, b) mentioned other grounds for processing personal data and c) referenced existing GDPR principles on automated decision making (see section "Interactions and overlaps with GDPR – consent" below in the part "General comments").Additionally, please find below some general comments on: • Explainability and notably on "Trust vs. explainability of algorithms": the technology does not (yet) exist to explain every AI decision. Furthermore, different algorithms, operational choices and business scenarios necessarily lead to different types / levels / expectations of appropriate "explainability". It is important to understand that AI models model a complex reality, so it cannot be expected that they "explain" complex reality in a scientific sense. Moreover, one cannot expect that a model perfectly "fitting" this complex reality could provide simple explanations that could be understood by everyone, including lay-persons. If the model describes perfectly a complex situation, a proper explanation of the decision mechanism will most likely be

complex as well.For example, economists have tried to explain financial markets or predict GDP, inflation and other economic aggregates for many years without ever succeeding completely due to the highly complex nature of underlying data. AI models strongly rely on the observation of dependencies among attributes which are too complex to be easily understood. It is possible to give an informative explanation of the model, the input data, the parameters, etc. However, it is not possible to explain the full process by which a specific decision has been made or reached. A requirement for a posteriori explainability would exclude most promising AI technique such as deep, recurrent neural networks and thus limit the ability to compete globally. Preventing the use of these technologies for purposes such as detecting financial crime, cyber security or terrorist financing is not acceptable since their use could lead to more accurate predictions.We thus suggest the following clarification as well as amendments: - We suggest distinguishing explicitly between a proximate explanation ("I got a drink of water because I felt thirsty") and a global, mechanistic one (the map from all sensory and hormonal inputs to all neuromuscular outputs of the brain). Once we move beyond simple trees, it becomes difficult to write a global, mechanistic explanation, even in symbolic language, that a human can understand. Proximate explanations are possible, while the more detailed global, mechanistic explanation often will not be.We should caveat that even proximate explanations can be rendered at different levels of precision and difficult to render at high precision. Local feature importance – often labelled as "explanation" – will sometimes not suffice even to describe a particular model output, for instance if higher-order, crossing dependencies are important. - Linked to the points above on explainability, the report suggests that, "to the extent possible, [we] design [our] system to enable tracing individual decisions to [our] various input". We suggest "possible and practical" instead of "possible," since a complete tracing would require recreating the model with features omitted, a task that becomes impractical as the feature count rises. We thus suggest the following amendment: "to the extent possible and practical, design your system to enable tracing individual decisions to your various input."AI HLEG draft ethics guidelines for Trustworthy AI, Page 23Ø - The guidelines also state that transparent "AI systems be auditable, comprehensible and intelligible by human beings". As already explained above, this principle of "intelligible by human beings" can severely limit the ability of AI to produce desirable outcomes should "human beings" not clearly be defined, so revising and limiting the expectation to systems being "auditable" would seem sufficient. Indeed, as mentioned above, one cannot expect all customers to be able to understand the ins and out of a specific algorithm. We thus suggest the following amendment:"AI systems be auditable".AI HLEG draft ethics guidelines for Trustworthy AI, Page 10 (third sentence under "The Principle of Explicability: "Operate transparently"")We need a spectrum/taxonomy of solutions that can be appropriate in different situations. These could range from high explainability (for

					business and regulatory reasons) to tested functionality – this is use case driven. For example, levels of explainability appropriate to different scenarios could include:§- Explaining the purpose but not the AI decisions to individual / customer;§- Providing a description of input data and optimisation factors to individual / customer on request;§- Providing short, automatically generated, description of an AI decision to individual / customer on request;§- Having a human analyse an individual / customer query and provide a response.Although the firm should have a good understanding of its own data processing, the appropriate level of detail provided to data subjects should vary, for example if there is a risk of "tipping off" a criminal and undermining crime detection systems.• Covert AI systems: although we agree that consumers should have a right to know if they are interacting with a machine, we would suggest amending this paragraph slightly and softening it in order to make a clear distinction between a right to know/ responsibility to disclose and an obligation to (re) inform consumers at each interaction. For some routine applications like queries for trade confirms, it would be overly cumbersome for clients to actively confirm awareness of interacting with an AI every time, creating a risk of "customer fatigue".		
Ninja	MARNAU	CISPA Helmholtz Center for Information Security	We welcome the European Commission's efforts to assess the transformative societal effects of Artificial Intelligence. The establishment of the High-Level Expert Group on Artificial Intelligence (AI HLEG) and its recommendations are the crucial starting point for the discussion on "Trustworthy AI made in Europe". We very much welcome the "Draft Ethics Guidelines for Trustworthy AI" (draft hereafter) and hope that the recommendations will fuel stakeholder engagement.While we welcome the draft, we would like to respectfully comment and discuss the draft's scope and results drawing on our expertise in the areas of information security and data protection regulation and technology.Introduction: Rationale and Foresight of the GuidelinesThe aim of the draft is to outline a human-centric approach to AI by ensuring the ethical purpose of AI as well as its technical reliability and robustness. We agree with and support these two main areas of focus for the document and appreciate that the AI HLEG sees the draft as a "living document that needs to be regularly updated over time to ensure continuous relevance". While the draft is intended to foster reflection and discussion, we are concerned that the document lacks incentives for stakeholders developing, deploying or using AI to practically apply the draft's recommendations. As all recommendations are referred to as voluntary and suggestions, the draft's impact is unclear.We understand	On page 5 the AI HLEG introduces a rights-based approach to AI ethics with the "additional benefit of limiting regulatory uncertainty". What would constitute this regulatory uncertainty is not clarified in the draft. We consider it of utmost importance to take stock of the existing regulation and drafts of upcoming regulation to assess and explain for which application fields and to what extent missing regulation or "regulatory uncertainty" actually exist.Also on page 5, the draft introduces the principle of autonomy as a core example of a fundamental right derived principle leading to informed consent as a value. While this is a helpful example to understand the relationship between fundamental rights, principles and values, in the context of AI it could be misunderstood as informed consent being the primary or the only legal base for AI use. In recent years with advancing digitalisation newer regulation such as the GDPR have acknowledged that informed consent (meaning knowing and understanding all consequences of data processing) has become less and less realistic in many circumstances of complex technologies. Art. 6 GDPR mentions informed consent as only one of several legal bases for data processing. Therefore, more paternalistic technology design regulation independent of consent has been introduced such as "Data Protection by Design" and the regulation of tracking technologies in the upcoming ePrivacy Regulation. Many AI	We strongly suggest to either add "Security" (meaning IT-Security) as an additional requirement or replace No. 8 "Robustness" with Security. From our point of view Robustness is a subcategory of IT-Security not the other way around. This is why the requirements under No. 8 miss crucial security requirements for AI: confidentiality and integrity (not only resilience to attacks), security against misuse by insiders, resilience and robustness against tempering with the learning process (adversarial learning), real-time guarantees, and intervenability. We would like to suggest that the section on "x-by-design" approaches on page 19 explicitly encourage the use of Privacy Enhancing Technologies and a data protection by design approach (mandatory under Art. 25 para. 1 GDPR).We think that it would be beneficial to extend No. 1 "Accountabilty" to also address accountability in complex distributed or federated processes.	The proposed questions are a useful first step for self-assessment. It could be beneficial to advance them into a framework for an "Ethics Impact Assessment".With regard to No. 8 Security/"Robustness" we would like to suggest to switch from asking about specific attacks to the more commonly used IT-security approach of defining the attacker model and its capabilities against whom you want to defend your system.	We thank the AI HLEG for the opportunity to participate in the Stakeholders' Consultation.While we very much welcome the "Draft Ethics Guidelines for Trustworthy AI", we are concerned that the draft gives a misleading impression with regard to the extent of existing regulation or legal uncertainty. We are also concerned that IT-security requirements that enable lawful and ethical use in the draft are limited to robustness and therefore miss crucial security requirements.We would like to encourage the AI HLEG to consult with experts in the field of EU data protection and technology regulation as well as IT-security experts to ensure that applicable legal requirements and the state of the art in secure system design are reflected in the guidelines and other upcoming documents.

that the AI HLEG will address questions of policymaking and potential regulation in its second draft (due in May 2019). While the draft states that respecting fundamental rights and complying with applicable regulation are a prerequisite for the AI's ethical purpose, the draft does not touch on the actual applicable regulation. The draft points out that "it should be noted that no legal vacuum currently exists, as Europe already has regulation in place that applies to AI", but unfortunately does not elaborate on what this regulation encompasses with regard to AI. Many of the non-committal guidelines in the draft are actual hard and enforceable legal requirements by European and national law. We are concerned that the draft might lead to the impression that the design, application and use of AI in Europe is mostly unregulated when that is far from the case. We respectfully suggest for the AI HLEG to consult with experts in the field of EU data protection and technology regulation to ensure that applicable legal requirements (such as Art. 22 GDPR) are reflected in the guidelines.

application fields are equally or more complex and, hence, may not lend themselves to use cases for the primary of informed consent. We ask that the AI HLEG would discuss the suitability of autonomy (e.g., by informed consent) and mandatory technology design for different application fields considering the already existing regulation.We appreciate the list of "families" of fundamental rights on page 7. However, we are surprised to not see Art. 7 ("Respect for private and family life") and Art. 8 ("Protection of personal data") of the Charter of Fundamental Rights of the European Union and Art. 8 ("Right to respect for private and family life") European Convention on Human Rights being mentioned as these have a crucial relation to data-driven AI technologies and are linked to several of the mentioned families of fundamental rights. The derived Ethical Principles (page 8-10) are certainly helpful principles for designing AI systems. However, we are concerned that they are too vague and open to interpretation by those designing and operating the AI systems to being able to introduce meaningful rights and protections for the individuals subject to these AI systems. This is why existing regulation should not be out of scope for the draft. It would be beneficial, if the draft would also offer any guidance on how to address conflicting human rights or principles when designing or using AI systems. The focus on utilitarian arguments of collective good seems to be excessive considering the jurisprudence of the European Court of Justice and the European Court of Human Rights.Critical concerns raised by AIIn 5.1 the draft raises concerns with regard to AI systems using biometric data. We would like to encourage the AI HLEG to reconsider the listed examples of biometric data use (listed are lie detection, micro expressions, voice profiling) as all of these face serious criticism from the scientific community as being not sufficiently based on evidence and scientific methodology.This raises another concern for AI systems that is not yet addressed in section 5. The use of AI decision making based on not scientifically proven assumptions (correlation vs causality) or pseudoscientific applications. Risks to ethical AI should not only encompass the risk of being unjustly identified but also the risk of being subject to unethical AI systems using not scientifically recognised assumptions or mathematical-statistical methods.We would also like to point out that targeted or mass surveillance for law enforcement purposes are not subject to the GDPR but to the Directive (EU) 2016/680 and its national implementation laws.We are unaware whether the AI HLEG will issue a report on technical guidance. But since the section 5 of the draft mentions anonymisation (it warns against insufficient de-identification) we would like to suggest to include information on Privacy Enhancing Technologies (such as Differential Privacy, Private Learning and Federated Learning) and encourage AI developers and users to consider more privacy-friendly designs for Machine Learning.We appreciate that the AI HLEG explicitly mentions increased risks in application scenarios with (informational, organisational, or legal) power asymmetries. This is why we would like to stress the need to discuss mandatory "contestability" in

				addition to transparency of AI decisions.
Catherine	Dupre	University of Exeter - UK	As a comparative constitutional law scholar, I have been working on the concept of human dignity in relation to human rights and democracy in Europe for the best part of the last 25 years. Of direct relevance to this consultation and these AI Ethics Guidelines, I recently published a monograph The Age of Dignity: Human Rights and Constitutionalism in Europe (Bloomsbury/Hart, 2015) and I am also the author of a detailed commentary on Article 1 EU Charter in S Peers et al (eds), The EU Charter of Fundamental Rights: A Commentary (Hart, 2014) and I am currently working towards the second edition. I regret that I have become aware of this consultation for too late to be able to engage with it more than very succinctly. My main comments stem from my understanding of human dignity in the EU.	It is indeed fundamental to draft these guidelines with reference to human dignity. The Draft Guidelines approach to this concept is however too limited in relation both to its actual development in the field of EU law and human rights, and to its significance for your guidelines. i) Human dignity not only promotes a human centric AI: it is both the most important value (first foundational value under Art.2 TEU) and first fundamental right (Art. 1 EU Charter), and its inviolable nature makes it stand out among all human rights. ii) The EU offers an exceptionally strong protection of HD and possibly the most detailed level of definition: see the entire title 1 and the mentions under Articles 25 and 31 EU Charter. Any assessment of HD needs to be done in full consideration of all these dimensions. The full title 1 of the EU Charter needs to be considered when thinking in terms of HD (not just Art.1 EU Charter). iii) HD captures a constitutional/human rights definition of humanity. The phrase itself is abstract and does not point to any particular dimension of humanity. It is the abstract nature of this concept that has made it possible to capture – and protect – an increasingly complex definition and understanding of humanity, reflecting social and scientific changes. It is therefore important to note that HD in the EU protects far more dimensions of humanity and far more human beings that the Draft Guidelines appear to consider. These include (as protected by EU law): - all human beings – not just citizens: HD protects all human beings equally regardless of nationality etc … - all human beings everywhere including in the work place (human beings as workers). This raises issues of working conditions, boundaries between robots and human workers, as well as the requirement to keep human benchmarks for assessing performance (i.e. not to expect that human beings can perform at the same level/in the

same way as AI).
- human beings lacking capacity to self-determine and consent (all those without autonomy): this is a key dimension and merit of HD. AI Guidelines need to consider all those who lack capacity and ensure that they are also protected. In particular, this is of direct relevance in AI uses in relation to patients with dementia. Protecting self-determination/consent etc… in interaction with AI is immensely important. This ought also to be considered for all those who have a limited (or none at all) mental capacity.
- protection of human beings beyond individual life span, consider protection of humanity over time, protection of future generations. AI uses today should also protect the human beings of tomorrow. This is particularly crucial in relation to ways in which AI affects human cognitive processes.
The human rights and constitutional definition/dimensions of human beings that is now protected through HD and by reference to its 'inviolability' has evolved from biological (born as a human beings), to genetic (e.g. prohibition of human reproductive cloning). Uses and developments of AI brings to the fore a cognitive dimension of human beings, that is not sufficiently captured by autonomy/consent/self-determination. Much of this is yet to be understood, but at the very least – in sort of precautionary manner – two principles might be tentatively considered for the purpose of these Guidelines. One is the retention of human cognition or cognitive mechanisms, skills and processes. The other is the human capacity of make mistakes and to make the wrong choices.

Carl	Schonander	Software & Information Industry Association	Full Comment Document on SIIA's Website: http://bit.ly/2Uv0AXm	Full Comment Document on SIIA's Website: http://bit.ly/2Uv0AXm	Full Comment Document on SIIA's Website: http://bit.ly/2Uv0AXm	Full Comment Document on SIIA's Website: http://bit.ly/2Uv0AXm	Full Comment Document on SIIA's Website: http://bit.ly/2Uv0AXmJanuary 31, 2019European Union High-Level Expert Group on Artificial IntelligenceRef: Stakeholders' Consultation on Draft AI Ethics Guidelines The Software & Information Industry Association (SIIA) appreciates the opportunity to comment on the draft ethics Artificial Intelligence (AI) guidelines. SIIA supports the discussion of such guidelines with the caveat that guidelines will not be uniformly applicable to all AI applications given that AI has such domain-specific applications. Defense, health, autonomous vehicles, marketing/advisor bots etc. each pose their own unique requirements. Even more broadly, SIIA considers that there should be a global alignment on a definition for AI developed with public and private sector stakeholders both to assist public policymakers and the private sector. Furthermore, SIIA notes there is a discussion about possible regulation of AI in the EU. SIIA reiterates that given how quickly technology develops in unanticipated ways, it is crucial for regulation not to focus on emerging technologies, i.e. regulation should be technology, a precept for which there is wide international support. Instead, regulations should be designed to prevent harm to consumers and businesses and crafted to address domain-specific situations, rather than how AI could be used in general. About SIIAThe Software & Information Industry Association (SIIA) is the principal trade association for the software and digital information industries. The more than 800 software companies, data and analytics firms, information service companies, and digital publishers that make up our membership serve nearly every segment of society including business, education, government, healthcare and consumers. As leaders in the global market for software and information products and services, they are drivers of innovation and economic strength – software alone contributes $425 billion to the U.S. economy and directly employs 2.5 million workers and supports millions of other jobs. For more information, please visit the SIIA Policy Home Page at www.siia.net. Introduction On September 17, 2017, SIIA released an Issue Brief entitled: "Ethical Principles for Artificial Intelligence and Data Analytics." The draft AI guidelines are consistent in many ways with what SIIA says in the Issue Brief. Our comments provide additional information on how disparate impact analysis studies could be conducted. This information is likely most pertinent to the profiling and law enforcement use case mentioned on page 28 of the "Working Document for stakeholders' consultation" and the Non-discrimination point on page 25 of the consultation document. Furthermore, SIIA concurs with the relevance of the ten elements described as "Requirements of Trustworthy AI" and offers additional comments on the Robustness (8) and Transparency (10) elements.Non-discrimination - Conduct Disparate Impact Analysis to Check for Bias With respect to the Expert Group's correct point in the Assessment List asking whether there "are processes in place to continuously test for such biases during development and usage of the system," SIIA considers that the way to address the possibility of bias is to conduct disparate impact tests as

appropriate. Note: in this context, "disparate impact" means an impact that has a disproportionate adverse effect on vulnerable populations. The principles guiding disparate impact tests reflect the widespread international norm that high-stakes decisions about people should not disadvantage vulnerable populations based on characteristics such as their race, gender, ethnicity, or religion. See the italicized text below from the SIIA Issue Brief for when and how to conduct disparate impact assessments. Since disparate impact occurs inadvertently, the only way an organization will discover on its own that its data practices have a disparate impact is to look for it. As noted above in the scope principle, organizations should put in place procedures and standards to determine when to conduct a full disparate impact assessment when they regularly develop, implement or use data analytic systems that might have a discriminatory effect on vulnerable groups. The following principles specify when a data analytic system should be subjected to a full disparate impact and what the elements of a disparate impact assessment are. • Organizations should evaluate a data analytic system for disparate impact when the design, implementation or use of that data analytic system has a significant potential for substantial and consequential discriminatory effects on vulnerable groups. • A disparate impact assessment determines whether a data analytic system has a substantial disproportionate adverse impact on a vulnerable group, examines whether the use of the system advances legitimate organizational objectives and compares it to alternative systems that might have a lesser disparate impact. Organizations regularly operating in areas that have consequential impacts on people's lives should evaluate data analytic system techniques for disparate impact when the design, implementation or use of data analytic systems has a significant potential for discriminatory effects. A disparate impact assessment has three steps. The first is to determine whether the data analytics system under review has a disproportionate adverse impact on a vulnerable group. This can be measured by standard statistical characteristics of the data analytic system such as departures from statistical parity or equal group error rates. Organizations should devise or adopt – in collaboration with academics, advocates, and independent technical experts – accurate and reliable guidelines and methodologies for detecting disparate impacts. The second step is examination of how the data system in question serves organizational objectives. Notwithstanding any disproportionate adverse effect on vulnerable groups, a data analytic system can pass a disparate impact assessment if it furthers a legitimate organizational interest. Avoiding disparate impact cannot be a requirement to abandon the values and goals that constitute an organizations mission. But furthering a legitimate objective is not sufficient to pass a disparate impact assessment, because there might be an alternative system that also furthers organizational objectives, but does so with a smaller impact on the vulnerable group. So, the third step in a disparate impact assessment is a comparison of the data system to alternatives. This step should involve an active search for alternatives to or

modifications of the system being reviewed. It should not be restricted to an assessment of obvious or readily available alternatives. Organizations should develop and assess alternatives to algorithms with a disparate impact to ascertain the extent to which they achieve organizational objectives. A data analytic system passes a disparate impact test, despite having a disproportionate adverse impact on a vulnerable group, when after an appropriate search for alternatives, an organization finds there is no alternative algorithm that furthers institutional objectives with a lesser impact. Disparate impact assessments should be conducted at the same frequency as other reviews needed to ensure the validity and reliability of models. Especially in the case of advanced analytic systems that improve in use, impact assessments need to be conducted frequently.It is crucial to emphasize the point above that the mere presence of a statistical disproportion involving protected classes is in no way a proof of legal liability for violation of non-discrimination laws. As noted above, these discrepancies are often an essential element in the use of algorithms to achieve legitimate business purposes. But they are an indication that further assessment is needed to determine the legitimate business interest served and whether there are alternative algorithms that could achieve the same result with less impact on the protected classes. For more detail on disparate impact assessments, see SIIA's Issue Brief on Algorithmic Fairness. Transparency - Communicate Key Factors in Scores and Evidence of Validity of Predictive Models SIIA notes that that the Assessment List's points 8 and 10 do not mandate disclosure of source code of proprietary algorithms, and SIIA considers this outcome correct. Companies need to be able to choose proprietary business models (or not) as they develop algorithms. Moreover, disclosure of such source code could allow bad actors to game analytical systems that defeat their purpose, like for instance criminals intent on credit card fraud. For SIIA's view on Transparency and Explanations, see the italicized text below from the Issue Brief. A key aspect of ethical use of data is an organization's willingness to be accountable to outside oversight about the processes and outcomes of data analytic systems. Accountability cannot be effective without transparency to the outside world and a commitment to conveying clearly and comprehensively how an organization's processes and standards address the ethical issues raised by data use, including how an organization assesses and remedies disparate impacts. Several U.S. and European regulations, described in the appendix on additional material, call for disclosures of explanations. The following principles regulate how an organization should approach these transparency questions. • Organizations should disclose what data they collect, the purposes for which it is used, and which analytic techniques and models are used to process data and produce an outcome. • Organizations should provide explanations of how advanced modeling techniques produce their results, including disclosing, where available and appropriate, the key factors that contribute to the outcome of an analytic process. • Organizations should publicly describe the model governance programs

they have in place to detect and remedy any possible discriminatory effects of the data and models they use, including the standards they use to determine whether and how to modify algorithms to be fairer. Trust in the fairness of a data analytic system relies on public awareness of data and the analytical systems used as well as the basis for organizational steps to detect and mitigate disparate impacts. Transparency about the process and standards used is especially important for disparate impact assessments, where ethical intuitions differ and social consensus on the right course of action might not be possible. The need to consult with public officials and the affected communities is especially strong in the cases, discussed below, of using sensitive variable in data analytic systems and determining how to navigate the tradeoff between accuracy and fairness when a data analytics system might not be able to fully satisfy both values. Organizations do not need to disclose source code of proprietary algorithms for several reasons. Disclosure is not useful for accountability purposes, especially in the case of advanced analytical techniques that improve themselves in use. Source code disclosure would likely produce counterproductive efforts to game analytical systems in ways that defeat their purpose. Disclosure would allow anyone to use or benefit from systems that require extensive development resources, thereby weakening the economic incentive in creating these systems. For these reasons, disclosure has not been required for heavily regulated traditional scoring systems such as credit scores that have been in use for decades. If organizations do not reveal their source code, they must take other steps to provide for transparency and accountability. Organizations should be prepared to communicate to outside parties the key factors that go into their scores, and to provide evidence on a regular basis of the continuing validity and reliability of the predictive models they use. Public trust in the fairness of algorithms requires sufficient disclosure so that people feel able to comprehend and assess the process used to produce insights that might have important effects on their lives. Need for Sector Specific GuidelinesRegarding the trustworthy requirement, the draft guidelines say "…in different application domains and industries, the specific context needs to be taken into account for further handling thereof…" They also specify that:"While the Guidelines' scope covers AI applications in general, it should be borne in mind that different situations raise different challenges. AI systems recommending songs to citizens do not raise the same sensitivities as AI systems recommending a critical medical treatment. Likewise, different opportunities and challenges arise from AI systems used in the context of business-to-consumer, business-to-business or public-to-citizen relationships, or – more generally – in different sectors or use cases. It is, therefore, explicitly acknowledged that a tailored approach is needed given AI's context-specificity." This is appropriate. The point the AI Study Group makes about regulation applies to ethics as well:"…attempts to regulate "AI" in general would be misguided, since there is no clear definition of AI (it isn't any one thing), and

							the risks and considerations are very different in different domains. Instead, policymakers should recognize that to varying degrees and over time, various industries will need distinct, appropriate, regulations that touch on software built using AI or incorporating AI in some way. ” The draft guidelines suggest that the final guidelines will emphasize this context-dependence by providing more specific guidelines for four distinct sectors. It might become clear in the discussion of these cases that the general guidelines are just elements to consider for appropriateness in a context, rather than requirements that must be implemented in all contexts.SIIA recommends that this point be articulated more clearly and completely in the final version of the guidelines. Relationship to Older Analytic TechniquesAs the guidelines make clear and others have as well, AI techniques, and particularly, machine learning programs are different and perhaps better ways of accomplishing the same tasks that earlier analytical techniques attempted to achieve. For instance, a machine learning credit score might do a better job of detecting when a person is a good credit risk than one based upon standard logistic regression techniques, but they are both attempting to do the same thing. Similar remarks apply to machine learning programs aimed at assessing the risk of recidivism, AI-powered data programs designed to improve the delivery of public services, content moderation algorithms, and facial recognition programs. The key thing is not the statistical technique used but the risks and challenges presented by the attempt to accomplish these tasks through data and data analysis.SIIA recommends that that the guidelines make it clear that the same ethical guidelines and regulatory rules apply to the application of analytics to achieve the same business or social objectives, regardless of the statistical techniques used. On behalf of SIIA, I would like to thank you for the opportunity to comment. Please do not hesitate to contact us if you believe we can be of further assistance. Sincerely,Carl SchonanderSenior Director, International Public PolicySoftware & Information Industry Association (SIIA)1090 Vermont Avenue, NWWashington, D.C. 20005United States
Joël	COLLOC	université du havre Normandie Normandy University France	None of PC hardware is currently developed in Europe. Europe is doing AI only in high layers of computer systems : the application layer. This is an important weakness, we are dependent of technology of USA and ASIA.Autonomous vehicle: There is a lot of responsibility problems (law) who is responsible of injuries to people when an accident occurs ? global health and wellbeing, climate change: Hell is paved with good intentions. To impose health measure and conditions of wellbeing against people willing, way of life and liberty of thinking is not ethic. Economical benefits generated by developing AI applications (usefull or not) is the main motivation of AI promotors. They try to convince people that AI is good for them and that the inconveniences are few confronted to the advantages. I agree that AI, like other computational application, must be governed by people for people and that the end-user must have the choice of using AI what to use, why, for what purpose,	The motto of France is « Liberté, égalité, fraternité » : « freedom, equality, brotherhood ». We think that the autonomous AI could alter the freedom, the equality by increasing the gap between the decision-makers and AI providers and the ordinary and vulnerable people. The dependence and the domination of the AI suppliers on the consumers, to whom they claim to bring benefactions, are going to make them to lose their freedom to decide on their life. Compliance with fundamental rights and applicable regulation is not enough because ethics is not the law and ethics is not a set of general rules, because what could be appropriate for some people will not be appropriate for other people (eg. religions, culture, way of life, political opinions). -Deontology charters or codes are often useless because nobody takes them into account when one want to gain money doing AI for business. -Ethics is not the law, it is a unique construction built for a specific	On AI ethics, other ethics problems rely on the medicine practice itself and the consequences of how the information is given to the patient in case of severe disease (medical ethics).Personalised medicine is based on predictive assomption based on genetic data that represent a probability of a disease to occur when may be this disease will never occur because environment, way of life and many other parameters are involved in the arisen of a disease and its evolution. The fact for a person to know what possible diseases could occur can bring her to a depression, anxiety and lead to the suicide. In the same way the disclosure of personal data to insurances, banks can make them to refuse a loan to this person in a very unfair manner.Ethics is not a mean to make AI technology to be accepted. Ethics is a prerequisite in the realization of any project of IA in particular when the system is autonomous and is taken to make decisions.alone.Bias : the main problem is	Ethics is not a mean to make AI technpology to be accepted. Ethics is a prerequisite in the realization of any project of IA in particular when the system is autonomous and is taken to make decisions.alone.To facilitate the auditability of AI systems is a late measure that only allows to understand that the system was inappropriate to the users. Ethics must be done previously of the system design to make sure that the AI system will fit the users's needsAccountability governance : the accountability have to give insurances that all users's needs and only user's needs are covered by the system and that if it is not the case what kind of penalty the providers will have to pay for.	Your definition of AI is not sufficient because it is only a list of features but is very confusing because it doesn't make a clear distinction between two main kind of AI:- the mimetic AI e.g. decision support systems to a user in specific domains using a knowledge base and a reasoning engine (eg. for diagnosis or diseases treatment) and -the Autonomous AI (like in robotics or autonomous cars) that intend to behave independently of a user and provides human being with services.The ethics problems are very different in these two kind of AI.

with what, when and where. That is the purpose of mimetic AI but not that of the autonomous AI. General principles (core principles and value) cannot ensuring ethical purpose because ethics is not the law but always a specific decision especially built for each person according to his way of life, culture, mind, psychology, trusts and mainly her/his personality and her/his willing and advices.Technically robust: To be technically robust is not sufficient because that only allows to verify that the system is running properly. But moreover the embedding knowledge of the system must be adequate and scientifically established in order to propose the appropriate advices to the user to help him to build a decision or to solve a problem. That is an obligation of mimetique AI applications. But autonomous AI must decide their choice by themself and may be against the willing and the opinion or choice of the human being end-user.To be technically robust is not sufficient because that only allows to verify that the system is running properly. But moreover the embedding knowledge of the system must be adequate and scientifically established in order to propose the appropriate advices to the user to help him to build a decision or to solve a problem. That is an obligation of mimetique AI applications. But autonomous AI must decide their choice by themself and may be against the willing and the opinion or choice of the human being end-user.AI is a powerful tool for powerful people that have the cognitive abilities to use these tools, the financial means to buy them or to develop projects to build them and that necessarily will lead to increase the gap between employers and employees and businesses and consumers at last between rich and poor people.The main negative impact of AI on people life is to suppress the work and then the mean of winning their life in a world where more and more tasks will be done by robots of any kind. What will be the way of winning a salary if you have not the cognitive ability to master the new AI tools. The society based on people at work must be replaced by new society based on sharing goods without working.Accountability depends of who is responsible of what AI devices are doing when they are doing wrong and cause damage to people.Data governance. A data is never independent of objects or people and thus when people are described by their data, they must have the opportunity to choose what is done with these data, what can be public and what must remain private. General principles of life is to manage an internal part protected from the external environment like a cell.Design for all, is impossible, because it is impossible to take into account all the advices and the needs of end-users if you don't know what is their way of life and what kind of help they want to use their tasks, duty and to solve their problems. A all purpose AI system is unfeasible.I don't trust that Autonomous AI could be really useful for humanity. We don't trust that Autonomous AI could be really useful for humanity. Only the mimétique AI (Decision support systems) can be adapted to end-users needs and way of life.Discrimination of AI is mandatory because the economical means of providers are much more important that those of ordinary people who will become subjects of these systems but not really willing end-users that decide what individual case in a specific situation never the same bu similar to previous cases and where people try to think what is the better decision to take in order to act with respect to the person(s) who is (are) concerned by the decision.

the diversity of human beings , cultures, way of life. Eg. The japanese son refused to visit his dying father because he knows that his father will not accept to see to have lost his honor in his son's eyes that he is ill and weak. In the same situation french people usually prefer to remain with the father to spend with him his last moments. This small example shows the complexity of cultural relativism and how a misbehavior could arise even if the intention was good.Training and education should be easy if and only if the AI system is really user friendly. In an Autonomous AI system the user is not really a user but a subject submitted to the system behaviour.

			is useful for them and what is useless for them but of value to the providers'benefits.Giving Information to stakeholders and Traceability of the AI system is a very good intention but is not praticable because the users will be buried in a big amount of informations and data that they will not have time to read and understand properly. Exactly like the e-document that is provided in software licences that nobody read and accept the contract without reading it because they have no time to do so and the unwilling acceptance is necessary if you want to use the software. If you refuse the conditions of the license you should give up to use it.It is important to sort AI tools that are suitable to useful needs freely chosen by the end-user and discard the others					
Alison	Hall	PHG Foundation	1. The PHG Foundation is a health-policy think-tank that has worked for over two decades on genetics and genomics. Its focus is on how novel, potentially disruptive technologies, can optimally be implemented into health systems to improve health care. The PHG Foundation is supportive of the aims of the High Level Expert Group on AI in formulating draft guidance that sets a benchmark for high standards which can be adopted throughout Europe. However, we have some concerns that this approach is predicated upon an exceptionalist view of AI. 2. Our experience of the regulation of genetic and genomic tests is that there are a lot of parallels between the proposed uses of AI and genomics: with both technologies – there is potential to generate potentially predictive and sensitive data which could be used in discriminatory ways. On the other hand, many genetic and genomic tests are uninformative, are routine, and do not yield sensitive data. Regulating all genetic/genomic tests on the basis that they are sensitive does not adequately distinguish between the different uses to which genetic/genomic tests might be put. 3. The same arguments can be made in relation to AI. Many applications of AI technologies pose no prospect of harm or benefit. We have some concerns that the tone of the ethics guidance is that AI is necessarily exceptional. We would like to see more consideration of the view that some applications of AI may be routine and may yield uninformative data. In such cases it might be neither proportionate or rationale to seek to impose an exceptionalist regulatory framework. There are, of course, some applications which require extreme levels of oversight; multidisciplinary expertise, and careful transparency. Mandating the same levels of oversight to all AI applications might risk burdening the sector with excessive levels of regulation.	4. The PHG Foundation acknowledges the importance of the long list of fundamental rights, principles and values that have been identified. However we suggest that there has been insufficient attention given to guiding developers and users as to how they should prioritise these principles when they conflict. By way of example, in medical ethics, medical research and the cases of withholding and withdrawing treatment are seen as paradigm examples of where harm could be caused to the individual, but where an action or omission is justified because it is mandated by other principles (such as respect for autonomy). It would be helpful for the guidance to include examples of where challenges might occur and advice as to how these potential conflicts might be resolved (e.g. the principle of non-maleficence: "Do no Harm", page 9). 5. Whilst we note that these guidelines are not intended to address legal and regulatory issues, there does however appear to be an assumption that informed consent will be the legal basis for data processing at numerous points. For example, in Chapter 1 there is reference to the need to obtain 'informed consent' and similarly a right for data subjects to opt-out of their data being processed through automated processing. We note that if data is processed through a legal basis other than consent (such as legitimate interest or public interest under Article 6 of the GDPR) and data is used for secondary purposes such as research, that there will not necessarily be an obligation to seek consent, so a right to opt-out will not be engaged. It would be helpful if these guidelines were to clarify how conflicts between these ethical and legal principles, such as the one described, might be reconciled with each other.	6. Section 4 refers to the need for an internal and external (ethical) expert. It is not clear whether the group are advocating for both an internal and external expert. Nor is it clear how independent that expert should be (and whether, like the Data Protection Officer under the GDPR) such a person is expected to have expertise in both AI and ethics. An independent ethical committee might be a more effective way of capturing expertise both in technological capacity and ethical issues – such a multidisciplinary group would be well equipped to advise on the ethical issues that might arise in response to a technological challenge (p. 8).	7. 'In view of AI's context-specificity, any assessment list must be tailored to the specific use case in which the AI system is being deployed'. (p28) The PHG Foundation has considerable expertise in assessing the impact of novel technologies for health services and health systems. We are engaged in active research which is exploring various aspects of AI use within healthcare, including for diagnosis and treatment (such as pathology, imaging) but also for screening for rare genetic diseases. Much of this work is available on our website at www.phgfoundation.org 8. Our view is that it would be better to consider the ethical and legal frameworks together than in isolation. A framework for proportionate and responsive regulation is already in place in the form of the EU General Data Protection Regulation, EU Medical Devices Regulations and EU Privacy Regulations. These are supplemented by industry standards such as IEC 82304 which operate across sectors (such as wellbeing apps and medical uses) [noted at p21]. These already take account of contextual issues such as the potential for interoperability and operating environment. 9. We think it might be premature for the high level ethics group to require developers and users to have additional 'ethics' expertise in the form of internal and external experts, before account is taken of the scope and impact of existing regulations (as is planned in phase 2 of activity of this group).	None provided

Anonymous	Anonymous	Anonymous	1. the document called " A Definition of AI: Main Capabilities and scientific disciplines" should make integral part of this document, especially of this introductory part. 2. the supplementation must include not only an approved and updated version of the AI's definition, but also multiple examples of the types of AIs because it is mandatory that the reader be able to identify as many of them and to have the representation of specific AIs that function in our days (examples from different domains, from marketing to stock market), in order to make the difference between different types of implications that one specific AI would have. From this point of view I consider that this introductory part misses the explanatory component which would help future readers to understand better the implications of AI's uses as well as the necessary values and principles that must be take into consideration in the future chapters. 3. The examples mentioned above should focus also on explaining the differences between software based AI and those hardware embedded as well as between those predictable and those unpredictable and less transparent which will be less or even non-transparent and impossible to inspect/audit. (neural network or genetic algorithms). 3. It would also be preferable that the examples should also identify possible negative effects on humans so that the next chapters of the document could be understood by taking into considerations future and possible risks not only the improvements that AI will bring into our life.	The list of values and principles must be supplemented by mentioning: -Predictability and - Incorruptibility			
Anonymous	Anonymous	Anonymous	We consider that this project is very pertinent and useful as a trustworthy AI remain a key element to allow this technology to be adopted and accepted by EU citizens. The "human centric" concept is a real "buzz" word but very vague. Shouldn't we stick to Human Rights. Among the objectives of a Trustworthy AI, no reference is made to peace and security, two fundamental elements of the European Union. p 2 When describing the two components of Thrustworthy AI, we would take out the reference to "an "ethical purpose""; which is very vague. The respect of rights, regulations, core principles and values is clear enough.	The Chapter 4 "Ethical principles" list a serie of principles that are too vagues ("Do Good"; "Do not Harm"; "preserve Human agency"; "Be Fair"...), too general and may even be in contradiction with some EU principles regarding the security and safety of EU citizens. These principles go far beyond the mission of the working group as they refer to concepts and values that should be discussed and determined by the EU political body (EU parliament and Commission). As indicated in the glossary, "Ethical purpose" should essentially refer to EU fundamental rights and applicable regulations while respecting EU principles and values. Full stop. And not going above these principles and values. The part 5 relates to some very points that are very questionable. The chapter on LAWS is very vague and point out a concept that is not properly defined neither well described through very basic and simple arguments. We consider this chapter may improperly stigmatize the role of AI in defense, at a time when defense and security issues are on the top of the EU agenda. Moreover, we have been extremely surprised that among these "critical concerns" no reference was made to "fake news", "opinion manipulation", and the use of AI to lure citizens via social medias and internet tools.	No comments	No comments	This document is very interesting and may be very useful for all the actors in AI. However, it should really stick to EU rules and principles and not trying to define a new sanctimony environment that goes beyond the EU commitment. Moreover, the absence of reference of peace and security issues is quite surprising as these subjects are at the heart of the EU agenda. The only reference to security and defense is to stigmatize a concept that is very "trendy" but do not refer to any technical concrete reality. Also, as underlined above, the absence of discussion around the use of AI related to the fake news issues and the opinion manipulation via social media is very surprising as this subject is at the heart of EU countries concerns. Lastly, if some non EU companies are well represented in this working group, we can regret that EU users and EU citizens do not appear to be represented.

				Objectively speaking, the impact of "fake news" and the use of AI and nudge science presents far more immediate risks for EU citizens, EU countries, European democracies and the EU itself than the potential development of LAWS in many years. The subject is very rapidly evocated in the document while it is currently one of the hottest issue in western democracies.	
Adam	Schlosser	Workday	Workday, a leading provider of enterprise cloud applications for finance and human resources, is pleased to provide input to the EU AI High Level Expert Group (HLEG) and strongly supports efforts to seek stakeholder input on this important issue. Workday's cloud-based applications are increasingly using AI and machine learning to empower enterprises to process a wide variety of HR and finance-related transactions, and gain new insights into their workforces and financial performance. We give customers real-time insights into their organizations, allowing them to make decisions grounded in data rather than guesswork. Being in the cloud also means that customers have access to their financial and workforce data whenever and wherever they need it. Our customers operate in environments that are highly complex and constantly evolving. Workday is headquartered in Pleasanton, California, with offices and customers across North America, Europe, and Asia-Pacific. We are deployed in over 200 countries and provide a user interface in more than 30 languages. Over 350 of our customers are headquartered in the European Union, representing a growth rate of 75% within the past year. These includes some of Europe's largest companies, such as Airbus, Sanofi, Siemens, and Unilever, and innovative and fast-growing companies like BlaBlaCar. Our European headquarters are in Dublin, where we employ a rapidly growing workforce of over 1,000 employees, focused on all aspects of our business from R&D, legal, sales, and operations. We believe the overall direction of the guidelines are a helpful first step and look forward to greater engagement with the HLEG and the Commission going forward. In addition to refining the current draft, we encourage the HLEG to create a plan to socialize the guidelines on a global scale in order to develop interoperable frameworks that will best demonstrate EU leadership on AI ethics. Singapore recently released a model AI governance framework and Japan has indicated the intent to make "human-centric" AI design a key pillar of the upcoming G20 meetings. Creating global best practices that reflect the EU's AI ethics design goals is a natural way to continue to put the EU at the forefront of AI innovation. While the guidelines provide solid context aimed at framing the importance and scope of the work, we think it is essential that the HLEG clarify AI and machine learning include many possible use cases - from automated actions, such as self-driving cars, to those use cases that streamline processes, such as spotting financial ledger anomalies. The definition specifically references activities that "decid[e] to take the best action," which seems to preclude a number of use cases	The Assessment List provided in Section III is a helpful first step towards implementing the guidelines. Workday believes every organization should be prepared to answer a similar list of questions and strive to create corresponding internal controls to operationalize their answers.However, because different AI use cases carry substantially different risk profiles, a threshold risk analysis should be added to aid organizations and policymakers in determining the depth of consideration needed in crafting answers to an assessment list.For example, some AI applications that objectively carry greater levels of risk to society, such as self-driving cars, will necessitate greater specificity of controls and corresponding documentation. Conversely, some activities, such as AI used to optimize a display on an application dashboard or automatically reading and processing vendor receipts, carry little to no risk and companies should not be expected to undergo the same extensive process when implementing the guidelines. There are also a number of use cases that fall in a gray zone between these examples and additional guidance on the questions to consider when determining risk will be helpful in efficiently implementing the guidelines. Creating a self-assessment tool will promote greater uptake of the guidelines because organizations and policymakers will gain additional confidence that resources are dedicated towards those uses cases where negative consequences are most likely to cause a loss of public confidence.At a minimum, an additional section on risk analysis should include questions aimed at identifying impacted groups and the type of potential effects (both positive and negative) with a goal of helping organizations create a proportional trustworthy AI implementation plan. Further discussion should occur on whether such a self-evaluation should identify specific harm categories (e.g. physical harms or potential violations of fundamental rights) or a broader classification system that might allow for more flexibility to self-identify types of risk and customize responses (or some combination of these approaches).We also suggest indicating that since products continuously evolve, so too should any risk analysis or internal controls. One possible approach is recommending something similar to penetration testing with cybersecurity, where processes are evaluated for potential negative consequences and remediation plans are transparently developed.Finally, as AI is heavily dependent on increased use, availability, and creation of data, Workday agrees strongly that a respect for privacy should be a foundational element of any AI ethics guidelines. To the extent any

			that serve mainly to provide better information for humans to take better actions and actual decisions are not automated.The extended definition in the separate annex does a very good job of getting into the nuances of AI, but these important nuances appear to be lost in the abbreviated version in the guidelines. Therefore, we suggest clarifying that AI may either "decid[e] the best action to take...or provide additional information to enable humans to take an action or make a decision."			additional details are envisioned related to data governance or data privacy, we would very much be interested in participating and sharing our industry leading strategy.	
Kirill	Tumanov	Maastricht University		* Page 9: The Principle of Autonomy is ill-defined. Stronger and more ellaborative wording is required.			* Page iii: Why only "European citizens"? Most of the AI technology is globally applicable and used. Or the European AI is built for Europe only? * Maybe should add: "To ensure AI explainability, AI components should perform operations which humans, of different levels of experience and expertise, should be able to replicate without use of the components, if such need occurs. Possibility of replication should be shown by the designers/developers of the AI components who propose them for use." That might be also useful for security of society at large.
Tomas	KLIEGR	The opinions expressed in this document are the author's own and do not necessarily reflect the view of his employer.	Issues raised are numbered. 1. The purpose of the document (who should follow the guidelines) could be made more narrow. The motto of the draft guidelines is "developing, deploying or using AI". What is the justification for including mix guidelines for users and developers of AI into one document? These are groups with very different backgrounds and possibly orthogonal objectives. What might be too technical for one group, will be in concrete for the other group group. 2. On conceptual level, in the introduction I lack emphasis on reaching the right balance between EU retaining competitiveness in AI and prospective regulation. (cf. "Europe is losing the AI race - The Washington Post") The document could be accompanied by a COST-benefit analysis, which would quantify the economic and societal impact of the measures proposed, both within EU and on exports of imports of AI products and services to other countries. 3. It would help clarity if the document gave early on a specific example of how the guidelines will: a) affect development of AI products, b) affect deployment of AI product, and c) affect user attitudes towards using AI products	4. The inclusion of explainability is very valuable, but I find the elaboration of the concept somewhat vague. 5. I am not sure it is right to adopt the new term "explicability", which - according to the document - was first published only in November 2018. While the notion of explicability is potentially very interesting and applicable, there has not yet been that much discussion and validation in follow-up research papers. The paper presenting explicability (cf. footnote 15 in the Draft guidelines) does not provide any empirical verification and provides only limited discussion of competing definitions, such as the the one presented in: Bibal, Adrien, and Benoît Frénay. "Interpretability of machine learning models and representations: an introduction." Proceedings on ESANN. 2016. There are multiple other papers dealing with explainability in various stages of the publication or review process available on arXiv, including those of the author. 6. The XAI section on page 21 provides a fragmented view of XAI purpose and challenges. The main message - in terms of impact on stakeholders - is formulated as follows: "it is necessary to be able to understand why it had a given behaviour and why it has provided a given interpretation." This formulation could be expanded to make it clearer a) who this applies to (algorithm developers or persons affected by the algorithm), and b) what qualifies as an explanation.		9. Missing clarification on the boundary between a "mere" algorithmic decision and an AI decision. This is important to determine that a specific process or algorithm is in the scope of the guidelines.	10. Definition of terms. The included Glossary is not sufficiently comprehensive. For the definitions introduced I lack references to scientific literature and other related established resources, such as Sammut, Claude, and Geoffrey I. Webb, eds. Encyclopedia of machine learning. Springer Science & Business Media, 2011. 11. The Draft guidelines are not sufficiently well aligned with a companion document "Definition of AI: main capabilities and scientific disciplines" available at https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december.pdf (further only "Definition of AI document") There is an unclear relation between the glossary included in the Draft guidelines and the complementary Definition of AI document. The two documents partly overlap, but sometimes provide colliding definitions. As an example of inconsistency, the Definition of AI document introduces terms that are not used at all in the Draft guidelines (such as "Narrow AI"), but it does not define key terms used, such as "explicability". Instead it defines the term "explainability", which is according to note 15 the Draft guidelines document superseded by "explicability" (see also point 5). 12. Throughout the document, there is a lack of emphasis on measurable aspects of AI. The document does not refer to specific methodologies that could be used for gauging aspects of AI affecting ethics or trustworthiness, such as the degree of

			7. When should explanation be provided. Currently, the Draft guidelines seem to indiscriminately recommend provision of explanation for all algorithmic (?) decisions. Is this feasible for software vendors to comply with? This could be aligned with the right to explanation in GDPR, which only - as far as I understand - affects certain types of algorithmic decisions. 8. What qualifies as an explanation. Can an explanation provided for a "black box" model by an "approximate" algorithm such as LIME or surrogate decision tree qualify as an explanation? For example, a surrogate decision tree may incorrectly explain a specific prediction. Should it be accepted as a valid explanation method - for regulators, for end users in a (non critical) recommender setting, for a life affecting decisions - like credit score or choice of medical treatment?		explainability of a particular machine learning model.
michael	eatwell	UNITE THE UNION GPM&IT SECTOR UK	- UNI Europa welcomes the possibility to contribute to the stakeholders' consultation and underlines the importance of a broad public debate and information on AI. This debate must result in clear ethical and social guidelines and standards with the aim of improving the living and working conditions of European citizens.- We acknowledge the innovative potential of AI and new technologies that can be beneficial for our society. However, these new technologies also create challenges and we are concerned about the possible risks and consequences relating to working conditions, skills and training, ethics, equality, health and safety (among others). Therefore, UNI Europa would like to underline the importance of addressing AI technologies and robotization as topics for collective bargaining at all levels (company, national and European). AI and robotics have a huge impact on the future labour market, as jobs will sometimes disappear or be transformed and other jobs will be created. We need to accompany this process and address the question of skills and training for the future workforce: need to ensure that training on necessary digital skills is provided by education institutions and companies, and that it is not the sole responsibility of the worker to keep up with the rapid technological developments. Employability needs to be promoted through upskilling and reskilling schemes for workers. Investment in formal, informal and life-long learning is key; we must enable people to work with AI or invest in competences that AI will not cover. It is important to develop action plans at EU and national level together with education providers and social partners in order to modernize education and vocational training. We therefore welcome the call from the ILO Global Commission on the Future of Work for "a formal recognition of a universal entitlement to lifelong learning and the establishment of an effective lifelong learning system. (https://www.ilo.org/global/topics/future-of-work/publications/WCMS_662410/lang--en/index.htm)- The social partners play a key role in this and the EU should cooperate with them and national governments in order to identify which job sectors will be affected by AI. We need to understand the	- UNI Europa supports the human-centric or human-in-command approach suggested in the guidelines. We agree that it is necessary that humans always need to remain in control of technology and machines. Likewise, we agree that the use of AI needs to respect European values and fundamental rights.- We recommend an expansion of 3.2 "In an AI context, freedom of the individual requires protection from direct or indirect coercion, surveillance, deception or manipulation." Much profiling that lies at the heart of AI systems relies on a degree of manipulation of data. This is not least relevant in relation to the world of work and especially in the use of AI in Human Resources. - We welcome that the HLEG understands the need to ensure that those involved in the development and marketing of AI (researchers, engineers, designers etc.) act in accordance with ethic and social responsibility criteria. This should be addressed by changing educational priorities for technical subjects and by providing lifelong learning opportunities (e.g. by incorporating ethics and the humanities into training courses in engineering). - Organisations and companies should develop tools to facilitate ethical discussions and decision-making throughout the whole design process. This should be completed by internal training programs on ethics for all employees. Such training should help employees understand the AI systems themselves, their rights in relation to said systems and their possibilities of redress, complaint etc. - AI should provide an opportunity for workers to apply their skills and competences to the fullest while at the same time remain owner of the production process. This includes the principle of transparency in the use of AI systems in HR, like the hiring of employees or the performance assessment of staff. It is important to safeguard the rights and freedoms of employees in line with non-discrimination rules as regards the processing of workers' data.- UNI Europa welcomes 5.1 – 5.4. We support that these examples raise real-life concerns of the adverse consequences of AI systems. - In 5.2. UNI Europa urges the group to expand on the issue of the human's right to know they are interacting with an AI identify. This	- In order to achieve "trustworthy AI", we need to establish public, independent and autonomous organisations that can control and audit (labour) algorithms (e.g. to identify underlying biases and the objectivity of data sets that train algorithms). Likewise, the implementation of the ethical guidelines on AI must be monitored. A European observatory focusing on the ethics in AI systems could play the role of an independent watchdog, including in business.- We would like the advice „to always keep record of the data that is fed to the AI systems" from the heading of data governance included under Accountability. For workers, it is paramount that the datasets used to evaluate performance, or in hiring or firing processes is transparent and can be accounted for. - The explanation of the principle of autonomy covers the question of AI at work only in a footnote, whereas this is an important issue that should be given a more prominent place. We would like to highlight the right of workers to individually and collectively opt out or withdraw from the use of AI systems (or a decision chosen by an AI system) if they undermine the workers' autonomy, decision making competence or disrespect fundamental rights and ethical principles. We recommend the inclusion of a special chapter that provides for ethical guidelines on AI in the work environment to address these issues more in detail.- UNI Europa welcomes that the HLEG on AI acknowledges the importance of social dialogue to realise trustworthy AI. We would like to add that the involvement of social partners, and in particular employee representatives, should not only take place regarding the general public debate on AI. Social partners should be involved in the establishment of codes of conducts, of standardisation schemes, development of training and in the proposed accountability governance. Employee participation and inclusion should take place early in the design, development and deployment of new technologies including AI and robotics. It is essential and important not only to inform and consult workers representatives in the work place or at branch level, but to enhance their co-determination rights and ensure their right to co-decide on the aims, reasons and	- We welcome that processes shall be examined in order "to allow a human control, if needed" (assessment list – governing AI autonomy). In so doing it should not be a question "to keep a human in the loop". We need clearly defined measures, that empower people to exert this control in all processes – regarding resources (technical equipment etc), organisational needs (time, liability, etc.) and qualification.- We suggest extending the list on the assessment of use cases (p.28) and add the question of processes, in order to use AI to ensure decent work (development and impact assessment).	- UNI Europa welcomes the call for Accountability Governance on page 21. The establishment of Data/AI Governance Councils in companies will indeed strengthen the accountability of AI systems and will address a weakness in the GDPR. The Council should consist of shop stewards and management and be responsible for holding management accountable and transparent to the use of AI and data. Whistleblowers should be able to address concerns to the Council and mandate the council to investigate on reported issues. - We welcome the process of developing guidelines for a trustworthy AI made in Europe, which encompasses corresponding "guidelines made in Europe", but would like to raise the question why non-European companies such as Google were granted full membership and full participatory right in the High- Level Expert Group. The status of associate expert would be more appropriate.- UNI Europa also supports the position of the ETUC regarding this consultation.

			timeline and extent of changes in the labour market. The involvement of social partners is a must to find appropriate and future-proof solutions to concerns relating to employment, training, the nature of work, (in)equality or social systems and collective bargaining, especially at sectoral level.- As AI and automation have the potential to transform not only simple tasks but very complex processes, we need to have a large public discussion about the areas in which the use of AI is reasonable and beneficial for society. Part of the debate should be the question of how the profits generated by AI should be re-invested for the common good by creating employment in domains such as care, health services, education or mobility. Employees should participate in the distribution of profits, e.g. through wage increases or reduction of working time. Moreover, AI wins should be used to strengthen social security systems. This could be a measure to address the problem of future job losses and the precarisation of employment relations in a-typical work (e.g. platform work) due to AI and automation.- It is therefore important to integrate the aspect of the quality of jobs, decent work and social progress into the ethical approach in order to create a balance with the purely economic objectives underlying the creation and use of AI and robotics. - The Human-centric approach (HCD) not only presupposes information, transparency, participation and traceability, but also requires specific negotiation processes regarding decision-making in view of the aims and implementation of AI-systems at a very early stage for stakeholders such as employees and their codetermination bodies.	could be done through a "labelling" system. For example, online bots should be labelled as such. Users should be made aware of the use of bots and AI in customer call-centre or help desks etc. - We would welcome that the employer-employee, employer-worker relation is explicitly mentioned in 5.3 as an example of power asymmetry. - Taking into account the power asymmetry in employer-worker relations, a separate point 3.6 on "workers' rights" should be added, which should contain the following points: "decent work by design", equal negotiation processes in the sense codetermination rights, informational self-determination of employees, non-discrimination principle and freedom of association including the right to strike. This is needed in order to secure worker's rights to co-decide on aims and application of AI systems, and create a legal framework.- Concerning the long-term risks and concerns we welcome that these should be considered. This could become an integral part of the accountability and audibility demands – i.e. that developers, users deployers etc need to reflect on the development/changing nature of the adopted AI as well as engage in predictions/forecasts of its future development scope and the consequences (positive and negative) hereof). - AI's influence does not only affect the world of work, but also democracy and society as a whole. We welcome that the draft refers to this point in Chapter I, paragraph 5.3., by stating that AI is not to be implemented in order to enable "citizen scoring" by a state/government. But this should also apply to private businesses. Neither states nor companies should be allowed nor have the possibility, to create human profiles such as "moral personality" or "ethical integrity". We reject the proposed opt-out-function and even possible "opt-in"-functions are not to be designed in a way that they conflict with fundamental human rights and possibly lead to the waiving of services that are useful for a person. AI-based services, that are important for work and life, must be designed in such a way that they do not require the collection of data which could be useable for human profiling.- Creating big data-bases always includes the risk of hackability as well as intentional and unintentional data-leaks. The guiding principle of "data-sovereignty" needs data-security in order to be viable. This implies explicitly not surveying data in areas that are of highly explosive nature for people in e.g. political, private or work-related areas. Fundamental rights as informational self-determination, the freedom of association and freedom of speech are not to be put at risk by creating such data-bases.	implementation of AI at the workplace. - Social partners at all levels should be involved in the implementation at company, industry, national or international level, including through collective agreements setting standards. In this context, it is required to describe the negotiation processes, e.g. central control structures for sector-specific solutions (cf. ‚AI Now' Report 2018): „Governments need to regulate AI by expanding the powers of sector-specific agencies to oversee, audit, and monitor these technologies by domain." (https://ainowinstitute.org/AI_Now_2018_Report.pdf)- Regarding the principles of accountability and transparency, we need to establish mechanisms for the protection of whistle-blowers who disclose the risks of AI systems or the non-respect of ethical principles – especially in the case of employees in companies that develop such systems. Internal reporting of risks and violations should be supported and rules in place to ensure follow up. - Organisations and companies should pay attention to potential biases encoded in the system development, training data and model performance – especially those that my affect the most vulnerable. They could also establish an internal ethical review process to democratise the decision-making process- Companies should not only increase transparency regarding the design and development of AI systems, but also in organisational chains of responsibility.		
Anonymous	Anonymous	Anonymous	I understand the definition of "Trustworthy" AI, but I miss any idea for reducing the necessary trust. Probably we could write about the so-called trust-less approach such as of the blockchain technology Bitcoin in this document. Especially now that blockchain technologies and networks such as Bitcoin are very popular and in many products are in relation with artificial intelligence.	Isn't every communication manipulation? I guess that the protection from manipulation, in the document, means protection from such manipulation that would do harm to the individual, but some type of manipulation, particularly that serves the human user, should be allowed, how else could the AI support a beneficial suggestion by arguments? If I am right, probably we should make clear in the text that what type of manipulation we would like to protect the individual from and what other type of manipulation we allow.	These guidelines are very useful when humans make the AI, but can we expand the same guidelines when an AI is created by another AI? There are fashionable methods already for writing programs that can generate other programs, see genetic programming. If we can create an AI, also we can create an AI that can create an AI. Thus, I suggest expanding the guidelines to AI-made AI-s in this document. Note where the guidelines are the same and where there are differences.	When discussing resilience to attack, I would note that, in fact, many tools that attackers use for malicious purposes were found to serve ethical purposes and many tools that attackers invented could be used in the right of a benevolent goal. Similarly, what to do when the same artificial intelligence could be used benevolently and maliciously, too? When should we develop an AI that might be used maliciously and what should we do when the malicious use-case turns up when the AI is in operation already?	While reading this working document I often found joy in wanting to make a comment about something and then finding this thing defined later in the text. For me, this experience is telling that the document follows good logic. Thank you, High-Level Expert Group, for your work so far!

Shahar	Avin	University of Cambridge, Centre for the Study of Existential Risk	-	-	"The list of "Requirements of Trustworthy AI" is a useful one. 'Robustness' and 'Safety' are particularly important requirements. They are both often individually mentioned in sets of AI principles, and there are extensive and distinct fields of study for each of them. Robustness is an important requirement because our AI systems must be secure and able to cope with errors. Safety is an important requirement as our AI systems must not harm users, resources or the environment. Robustness and safety are crucial requirements for trustworthiness. As an analogy, consider that we could not call a bridge 'trustworthy' if it was not reliable and resilient to attack, and also safe for its users and the environment. These two requirements are importantly distinct from the other requirements, and work best as stand-alone requirements."-The report "invite[s] stakeholders partaking in the consultation of the Draft Guidelines to share their thoughts on additional technical or non-technical methods that can be considered in order to address the requirements of Trustworthy AI." We would like to share some additional technical and non-technical methods that are not yet on the list. These are mostly drawn from the major Febuary 2018 report The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. We co-authored this report with 26 international experts from academia and industry to assess how criminals, terrorists and rogue states could maliciously use AI over the next five years, and how these misuses might be prevented and mitigated. When released this report was covered across Europe and welcomed by experts in different domains, such as AI policy, cybersecurity, and machine learning. We have subsequently consulted several European governments, companies and civil society groups on the recommendations of this report. The European Union's Coordinated Plan on Artificial Intelligence, published on the 7th of December 2018, mentions the importance of the security-related AI applications and preventing malicious use.Several of the methods we explored are already mentioned in the Guidelines, such as codes of conduct, education and societal dialogue. However we also explored some methods that you do not yet mention. Our report made recommendations in four 'priority research areas'. In this response we split these into 'technical' and 'non-technical' methods.• Learning from and with the Cybersecurity Community• Exploring Different Openness Models• Promoting a Culture of Responsibility• Developing Technological and Policy Solutions**Technical methods** *Learning from and with the Cybersecurity Community*Formal verification. The use of mathematical methods to offer formal proofs that a system will operate as intended. In recent years this has worked on complex systems, including the CompCert compiler and the seL4 microkernel. It could be applied to AI systems. Security tools. Software development and deployment tools now include an array of security-related capabilities (testing, fuzzing, anomaly detection, etc.). Tools could be developed to make it standard to test and improve the security of AI components during development and deployment. Tools could include: automatic generation of adversarial data; tools for analysing classification errors;	-	This response was written by Shahar Avin and Haydn Belfield from the University of Cambridge's Centre for the Study of Existential Risk, a research group which studies the security implications of emerging technologies. For the last five years we have been closely involved with the European and international debate about the ethical and societal implications of artificial intelligence (AI). These Draft Ethics Guidelines are an important, concrete step forward in the international debate on AI ethics. In particular the list of technical and non-technical methods and the assessment list will be useful to researchers and technology company employees who want to ensure that the AI systems they are busy developing and deploying are trustworthy.

automatic detection of attempts at remote model extraction or remote vulnerability scanning; and automatic suggestions for improving model robustness.Secure hardware. Increasingly, AI systems are trained and run on hardware that is semi-specialized (e.g. GPUs) or fully specialized (e.g. TPUs). Security features could be incorporated into AI-specific hardware to, for example, prevent copying, restrict access, and facilitate activity audits. *Exploring Different Openness Models*Central access licensing models. In this emerging commercial structure, customers use services (like sentiment analysis or image recognition) from a central provider without having access to the technical details of the system. This model could provide widespread use of a given capability while reducing malicious use by, for example: limiting the speed of use, preventing some large-scale harmful applications; and explicitly prohibiting malicious use in the terms and conditions, allowing clear legal recourse. *Promoting a Culture of Responsibility*Differentially private machine learning algorithms. These combine their training data with noise to maintain privacy while minimizing effects on performance. There is increasing research on this technological tool for preserving user data privacy.Secure multi-party computation. MPC refers to protocols that allow multiple parties to jointly compute functions, while keeping each party's input to the function private. This makes it possible to train machine learning systems on sensitive data without significantly compromising privacy. For example, medical researchers could train a system on confidential patient records by engaging in an MPC protocol with the hospital that possesses them. Coordinated use of AI for public-good security. AI-based defensive security measures could be developed and distributed widely to nudge the offense-defense balance in the direction of defense. For example, AI systems could be used to refactor existing code bases or new software to security best practices.Monitoring of AI-relevant resources. Monitoring regimes are well-established in the context of other dual-use technologies, most notably the monitoring of fissile materials and chemical production facilities. Under certain circumstances it might be feasible and appropriate to monitor inputs to AI technologies such as hardware, talent, code, and data.**Non-technical methods***Learning from and with the Cybersecurity Community*Red teaming. A common tool in cybersecurity and military practice, where a "red team" composed of security experts deliberately plans and carries out attacks against the systems and practices of the organization (with some limitations to prevent lasting damage), with an optional "blue team" responding to these attacks. Extensive use of red teaming to discover and fix potential security vulnerabilities and safety issues could be a priority of AI developers, especially in critical systems.Responsible disclosure of AI vulnerabilities. In the cybersecurity community, "0-days" are software vulnerabilities that have not been made publicly known, so defenders have "zero days" to prepare for an attack making use of them. It is common practice to disclose these vulnerabilities to affected parties before publishing widely about them, in

					order to provide an opportunity for a patch to be developed. AI-specific procedures could be established for confidential reporting of security vulnerabilities, potential adversarial inputs, and other types of exploits discovered in AI systems.Forecasting security-relevant capabilities. “White-hat” (or socially-minded) efforts to predict how AI advances will enable more effective cyberattacks could allow for more effective preparations by defenders. More rigorous tracking of AI progress and proliferation would also help defensive preparations. *Exploring Different Openness Models*Pre-publication risk assessment in technical areas of special concern. In other dual-use areas, such as biotechnology and computer security, the norm is to analyse the particular risks (or lack thereof) of a particular capability if it became widely available, and decide on that basis whether, and to what extent, to publish it. AI developers could carry out some kind of risk assessment to determine what level of openness is appropriate for some types of AI research results, such as work specifically related to digital security, adversarial machine learning, or critical systems. Sharing regimes that favour safety and security. Companies currently share information about cyber-attacks amongst themselves through Information Sharing and Analysis Centers (ISACs) and Information Sharing and Analysis Organizations (ISAOs). Analogous arrangements could be made for some types of AI research results to be selectively shared among a predetermined set of ‘trusted parties’ that meet certain criteria, such as effective information security and adherence to ethical norms. For example, certain forms of offensive cybersecurity research that leverage AI could be shared between trusted organizations for vulnerability discovery purposes, but would be harmful if more widely distributed. *Promoting a Culture of Responsibility*Whistleblowing measures. Whistleblowing is when an employee passes on potentially concerning information to an outside source. Whistleblowing protections might useful in preventing AI-related misuse risks.Nuanced narratives. There should be nuanced, succinct and compelling narratives of AI research and its impacts that balance optimism about its vast potential with a level-headed recognition of its challenges. Existing narratives like the dystopian “robot apocalypse” trope and the utopian “automation boon” trope both have obvious shortcomings. A narrative like “dual-use” might be more productive.		

Matthias	SPIELKAMP	AlgorithmWatch together with members of the ELSI Task Force of the Swiss National Research Programme 75 on Big Data	Major comments:The specific role of trustworthiness as the focus of ethical guidelines should be clarified:1) what makes AI trustworthy in addition to 'reliable', or 'ethical'? What is the relation between these concepts? What are the differences between them? What is the advantage of having a code for 'trustworthy AI' rather than an ordinary ethical code?2) is AI supposed to be trustworthy or the people behind it, or a combination of both?Furthermore, is there only a single relation of trust, or different ones? Are the same recommendations relevant for trust between AI systems and the computer scientist that create and maintain the system and trust between an AI system (including the humans responsible for its maintenance) and its users?More generally, it would greatly help the guidance given to be provided a list of potential candidates for trustworthiness. The document remains very vague on who is to trust and who, or what, is to be trusted. IT is not even clear what are the implied 'identity criteria' for the AI systems that are supposed to be regulated. For example, if we consider AIs in which algorithms fed on data are embedded (as in many services based on machine learning algorithms), then one could ask how updates of the AI affect its identity and, a posteriori, the trust relationship with humans (users, engineers, controlling agencies etc.) that has been established and possibly nurtured so far. At what stage of any update does the original AI stop to be 'itself' and mutates in something different and distinct from the original? Is the re-training of a machine learning algorithm (even in case of a single modification of the original training dataset) enough to trigger an identity shift? Should we reconsider the trust relationship so far, or is the AI still 'itself'?Finally, the document fails to highlight what is truly special about trust-based relations and a trust-based society. While it emphasizes transparency both at the level of fundamental principles and in practices, there seems to be no realization that transparency may not be at the centre of trust-based relations. Indeed, one can argue that one of the distinguishing characteristics of placing trust in others is precisely the willingness to rely on a third party without the ability, or even the need, to check what the other party does. This is not to say that transparency is useless in a trust-based society. But transparency appears to play a complementary role: while most people use AI because they trust it, few people are expected to be inquisitive if there is trust. On the definition of AI: the definition of AI describes AI as a system that acts in the physical or digital world. But many potential software applications that these guidelines seem to intend to address are not artificial agents. For example, statistical models that provide assistance to human decision makers, without substituting them, do not act in the physical or digital world. Is the guideline not intended to address the concerns raised by those models? Or if so, should the definition of AI be revised? "Bias is a prejudice for or against something or somebody, that may result in unfair decisions" (iv): it may be worth stressing that statistical bias will be intrinsic to decisions based on statistical predictions in a context in which features of interest are not distributed homogeneously in different sub-groups of the population (e.g. men and	The term 'values' is used here to identify more concrete entities than principles and rights, for example 'informed consent' is described as a value. This is unusual for both philosophical ethics, where values identify broader and more general concerns, such as equality, efficiency, freedom, etc., and everyday language (do lay people really think of 'informed consent' as a value?). • Is this list sufficient to establish an ethical purpose? What about when core values differ between member states?• The human-centric emphasis discriminates against non-human animals. NHAs represent another vulnerable group and AI should respect them too, even if humans are more important.Understandably, the guidance is not committed to one ethical framework in particular and does not provide a decision rule to resolve potential conflicts and trade-offs which may arise. Any such framework would certainly be considered more controversial than a list of principles and rights to be weighted against each other in a context-sensitive way. However, it is to be expected that conflicts and trade-offs arise at the level of principles, rights and values. Thus, the guidance could be improved by providing some indication of procedures for assessing trade-offs. The draft wants to establish elaborated monitoring and assessment routines for AI, which are aimed at the public discussion and should ensure public 'trust' in AI systems. Therefore, the five guiding principles provided are being considered as the main normative basis of judgements of AI trustworthiness and complemented by rights, values and checklists, at different levels of abstraction. What seems to be missing is an indication of some procedure to attach weights to the different principles and solve disagreements when people disagree on which principle should have priority in a given situation. Or at least, the limitations of an ethical framework that at the most fundamental level relies on prima-facie ethical principles to be traded-off against each other intuitively could be explicitly acknowledged and the need to develop forms of ethical deliberation to solve these trade-offs could be mentioned. On social scoring (12). The paragraph moves abruptly from 'normative citizen scoring' concerning 'all aspects and on large scale' to scoring in limited social domains. The document recognizes that scoring in a limited social domain refers in some cases to established social practices, including practices such as education and driving licenses, that are commonly accepted, at least when AI is not used. But what is meant here by 'normative citizen scoring' is entirely unclear: an example seems to be needed. Concerning the opt-out option for domain-specific social scoring, why should there be an opt-out option when AI is used but not when AI is not used? E.g. should statistical models used to identify tax evasion be handled differently - providing an-opt out option - when AI is used, but not otherwise? What does it mean to have an opt-out option? Does it imply a right to be judged without the use of the AI? And what does that mean? To be judged by a human without the use of knowledge from statistical models? If not, why do traditional statistical models (e.g. actuarial tables in insurance) differ from AI in terms of implying a right to opt out from scoring? Otherwise, suppose that the right to opt out from social scoring	Major comments:One methodological problem with this section seems to be that it fails to examine the matter at hand at a coherent level of abstraction. For example, at p.15, the text includes a recommendation to ensure that data from the same person does not end up in the training and test set. This is a very specific recommendation about a specific way to mitigate statistical bias, but why does it deserve such special status? Moreover, the section makes no mention of the trade-offs that may arise from the need to implement different rights, principles and values. For example, the recommendation to keep track of all data fed to the AI system (p.15 sect. II) and the similar claim that AI "systems should document both the decisions they make and the whole process that yielded the decisions, to make decisions traceable" (p. 20) may be in trade off with the requirement to protect the privacy of the persons affected by the decision. This demand could, further, be in conflict with intellectual property rights and security disclosures. Models and systems are often considered a trade or governmental secret. How should this be regulated if the value transparency and intellectual rights or security are at odds with each other? How should these processes be implemented and how can it be ensured these rules of traceability and auditability are not violated?Similarly, "the capability to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environments, as well as the provenance and dynamics of the data that is used and created by the system" will tend to deliver a system that is transparent and potentially vulnerable to be manipulation. Transparency is sometimes alleged to conflict with protecting an AI decision system from malicious or self-serving attempts to manipulate their outcomes by strategically responding to them, which may lead to unfairness (e.g. between persons with different degrees of understanding of the logic behind the algorithm). We are not claiming that it is always socially desirable that the logic of algorithm should be kept opaque. But since this is an objection that is sometimes raised, by stakeholders and even regulators, against the demand of more algorithmic transparency, it would be useful if the working group were able to provide some advice on the matter to future regulators. The section (5) on non-discrimination (p. 16) should stress that discrimination does not necessarily derive from the data (e.g. biased social practices producing the data or incomplete data) but is, in a certain sense, a non-avoidable feature of all decisions grounded in statistical predictions (which typically are only imperfectly accurate, and even if perfectly accurate, may still be objectionable). Bias can also arise from data that perfectly represent the 'ground truth'. This is because a model may appear discriminatory if it treats individuals in different groups in very different ways, even if there is no bias in the data and the data somehow 'justify' this. For example, suppose that data from online learning platforms truthfully report that women are less likely than men to select a STEM subject when given the choice. Even if the statistics maintains external validity over time, society may reasonably object to a recommendation algorithm that recommends	It will be a challenge to make these recommendations content-dependent. It would be helpful if, also outside this document, examples and illustrations would soon be produced.	2. Special status of AI? Should AI be up to special ethical standards to be trustworthy, or the same standards as non-AI-involving social practices in the same domain and fulfilling the same function as AI? For example, social scoring is an old and established social practice, even before AI. As the authors recognize, driving licences and grades at school are forms of social scoring. General questions: are the general principles used to specify goals of trustworthy AI analogous to those of not AI-based practices? If not, why should they differ?Written and endorsed by:- AlgorithmWatch (Sebastian Gießler and Matthias Spielkamp)- Andrea Ferrario, Scientific Director of the Mobiliar Lab for Analytics at ETH; Department of Management, Technology, and Economics, ETH Zurich- Members of the ELSI Task Force of the Swiss National Research Programme 75 on Big Data-- Michele Loi, Digital Society Initiative and Institute for Biomedical Ethics and the History of Medicine, University of Zurich (Lead writer)-- Markus Christen, Institute for Biomedical Ethics and the History of Medicine and Digital Society Initiative, University of Zurich-- David Shaw, Institute for Biomedical Ethics, University of Basel-- Christophe Schneble, Institute for Biomedical Ethics, University of Basel

women being unequally likely to be liable for road accidents and to be involved in violent crimes after release from prison) and that the use of statistical criteria as a basis of decision making can be controversial - depending on the context - irrespective of issues of accuracy and bias, especially when the role for human judgment is limited or absent altogether. And among prejudices, one could perhaps also mention people's experiences (including education, cultural and religious background) as source of bias, for humans and models trained on human data as well. Generally, it may be worth stressing that 'getting rid of bias' is not a sensible and feasible policy goal. Instead, policy requires making deliberate, reasoned, if possible principled and publicly legitimated choices concerning which biases to accept and which to mitigate or neutralize when optimizing models. The unavoidability of some form of bias/discrimination/unfairness in decision making that relies on statistical predictions results from the 'fairness trade-offs' between different definitions of bias and discrimination, highlighted by the computer science literature of the past few years.Minor comments:"Given that, on the whole, AI's benefits outweigh its risks": this is a sweeping statement. What is the evidence on which it is based? Was a cost-benefit assessment of AI technology as a whole conducted? "It can help achieving the sustainable development goals such as promoting gender balance" (1): in what sense is gender 'balance' a goal of sustainable development? Official UN documents talk about gender equality that is both a different and arguably broader political goal than achieving some kind of arithmetic balance (e.g. 50-50%, or less?) in representative organs or education.Non-discrimination is described in terms of the same rules applying for everyone to access a list of goods (7). This is an odd definition of non-discrimination because the 'same rules' that apply to everyone may be discriminatory rules, e.g. rules designed to exclude certain types of people from the goods (or achieving this unintentionally). The problem here is not that the 'same rules' are not applied, but that the rules are unfair and discriminatory.The right to be informed of any automated treatment and to be offered to opt out (p. 7) is too general and thus implausible. Should a person with precedents for crime have the right to be informed of an automated treatment by an algorithm of the police that is used to narrow down the possible suspects of a new crime, when the procedure is implemented legally? Should the person have a right to opt out of this treatment? On trust and AI: at p. 8 it says `Trust is a prerequisite for people and societies to develop, deploy and use Artificial Intelligence'. We disagree with the statement. At the current stage, ignorance about services backed-up by AIs is commonly widespread; on the other hand, AI-based services are increasing in number in different sectors. Typically, users access those services without being aware of the presence of AIs, or in absence of alternatives. As mentioned above already, there is no trust building dynamics in absence of a clear identification of the trustee, or in presence of constraints on the trust or decision-making. Therefore, the above statement does not reflect the status quo, and could be amended in specifying by AI implies a right to opt out from any social scoring (irrespective of AI is used). If so, should the person who opts out bear the cost of not being socially scored? These costs can be considerable, as they may include being unable to obtain credit (in the absence of a creditworthiness score), being unable to drive (in the absence of a driving licence), and being unable to obtain an education (in the absence of grades). If, finally, the person who opts out from social scoring should not pay the price of her opting out decision, how could practices that rely on social scoring to be sustainable be guaranteed? E.g. email relies on the scoring of email senders to activate spam filters: should a spammer have the right to opt out from this scoring and yet be allowed to send spam around? On section 5:5.1 – Identification without consentInterestingly, 'identification without consent' does not refer to companies identifying individuals without asking their consent. Rather, this section addresses the possibility that, even if companies ask and obtain the (formally) informed consent of individuals, the informed consent provided online by citizens should not be taken at face value. This section of the document contains one of the strongest statements in any public document so far about the inadequacy of the 'notify and consent' strategy for dealing with privacy/ data protection. A strategy that is, and has been for decades, the main procedural solution to achieve privacy and autonomy without sacrificing either. The authors write that, in the light of the fact that 'consumers give consent without consideration', there is"an ethical obligation to develop entirely new and practical means by which citizens can give verified consent to being automatically identified by AI or equivalent technologies."We also believe that the system of privacy/data protection revolving around the current version of online informed consent as its main pillar is largely inadequate, at least for high stake decisions based on personal data. Yet this much needed critical section raises more doubts and puzzles than it solves, in the context of this document:1. There seems to be a contradiction between scepticism about informed consent as a procedural solution of difficult governance issues and elevating 'informed consent' to the status of a value.2. While it is undeniable that online informed consent procedures are sufficient to legitimize the uses of identifying technologies, it is not so clear what could substitute it in the context of AI used for online services. The section criticizes informed consent as inadequate in this context but, for lack of alternatives, leaves no option for identification technologies, outside the extreme one that identification technologies cannot be justified on the basis of a preference or desire of the consumer, until radically new forms of consent (of what kind?) will be developed. The only exception are goals (such as detecting fraud, or terrorist financing) where the justification of re-identification is independent from the informed consent of the subject of surveillance.3. The strong claim that 'consumers give consent without consideration' raises the problem of informed consent as an instrument of legitimation in general for AI, not only in relation to the specific identification technologies in question in this section of the document. If STEM subjects more often to men than to women. Hence, it seems important to introduce and stress the idea that the goal of 'avoiding discrimination' only makes sense relative to a prior value judgment about the kinds of inequalities that society deems permissible, even desirable, and those that are considered unjustifiable. The report could stress the importance of promoting a wide societal debate about the nature of bias, unfairness, and discrimination with statistical predictions, that attempts to reconcile conceptualizations from common-sense, ethics, law and statistics. We invite the Expert Group to provide recommendations on advancing a more transparent and informed debate about the fairness metrics that have been proposed in the field of computer science, which appears to be crucial for their political legitimation. This may include the promotion of policies that advance the public understanding of the different forms of unfairness and discrimination that may arise through the use of AIs and, more broadly, statistical models (also, already in use) both in high-stake decisions and low-stake decisions with serious cumulative effects.It is good that the AI HLEG is recognising the specific theoretical and methodological characteristics of AI development. The question of accountability, and the shift of this accountability to the user, no matter if the system is a 'black box' or not, could be an important guideline for AI development and regulation. The epistemic (methodological) values of traceability and auditability could, however, be at odds with the epistemic and scientific features of AI development. AI development and research are heavily influenced by the epistemic cultures of the disciplines informatics and computer science. Most branches of computer science are concerned with 'making things', like computers, algorithms or software which should solve a specific problem for governments or businesses. A possible problem, however, is that the instrument of accountability is pointless if society lacks persons with the information, skills, motivation and time to assess the achievement of the relevant desiderata by AI systems. Some information will only be distributed within the companies and even the skills necessary to make sense of information made public are very unequally distributed in the population. The majority of the population hopes to be able to 'trust' AI. But trust is only well placed if a more competent, motivated, inquisitive, and sceptical minority is able to assess if such trust is well placed. Recognizing the relation of dependency between an expert community and the broad population should lead to:• stress the importance of making information about AIs available even if it is not understandable by the average users of AI, • invoke measures that facilitate the acquisition of relevant information and skills by consumer groups, trade unions, and other groups representing stakeholders of AI;• stress the importance of legislation protecting whistleblowers, as the public has otherwise no access to the inner working of the AIs that are not made public. History shows that without whistleblowers it will in fact be very difficult, if not altogether impossible, to determine violations of privacy taking place within companies. (The issue of whistleblower protection is not even

that it is about attributes on the AIs that we, as society, would like to be developed.Complementary set: what are examples of NOT trustworthy AIs? Which are the legal consequences of having them either offline or online in IT systems of companies or agencies? Is there the possibility of drawing an analogy with the personal use of drugs?- "To avoid harm, data collected and used for training of AI algorithms must be done in a way that avoids discrimination, manipulation, or negative profiling" (p.9): is this plausible? Is it not more helpful to provide criteria when negative profiling is legitimate and where it is not? Is the question one of avoiding negative profiling or of societal control of the legitimacy of profiling, especially when negative? Should positive profiling be considered equally problematic, given that it can lead to denying advantages to some people, causing inequalities which may be unjustified?Listing "respect for human dignity" as a fundamental right is not really helpful as ethical guidance, it's a vague term (but popular at the EU level) that does not really 'unify' or help to assess trade offs between its listed components: 1) humans' physical and moral integrity, 2) personal and cultural sense of identity and 3) the satisfaction of their essential needs. Why not including the three elements of human dignity as distinct principles?- The principle of justice (p.10)This is an extremely important principle, but its discussion appears to be incomplete in two different ways, concerning respectively the aspect of fairness/discrimination in statistical prediction and justice in the utilization of data resources. Concerning the first, the guideline prescribes that "positives and negatives resulting from AI be evenly distributed". This is very unclear. The language of positives and negatives seems to refer to the context of statistical prediction. If so, first, it is unclear who are the subjects of distributive justice: legally 'protected groups'? Vulnerable populations? (The two are not the same). Second, it is unclear what 'evenly distributed' means, for example, if different groups have different baseline distributions of the predicted attribute (e.g. violent reoffending for prisoners released on parole) should 'evenly distributed' entail that women and men should have the same probability of being released; or should it mean that women and men who do not reoffend should have the same probability of being released, etc.? There is also an issue of trade-offs with the requirement of avoiding bias since it is mathematically proven that enforcing both aforementioned 'even distribution' criteria comes at the expense of predictions being unbiased in a different sense (e.g. equally likely to be correct for the different groups). The document lacks any reference to the often discussed question of trade-offs between fairness objectives/metrics and fails to discuss what a possible role of future public institutions could be, with respect to providing guidance on how to resolve these 'hard questions' of machine learning fairness (that have been sometimes referred to as the "trolley problem for machine learning". (p.15 (section on discrimination) mentions that data always carry some sort of bias; but the bias may not be in the data, but rather in the inferences drawn from that. See below our commentary on that section.)Concerning informed consent is not to be relied upon as a legitimation mechanism, because it is always given 'without consideration' this leaves a huge regulatory void, as informed consent is one of the cornerstones, if not the most important cornerstone, of the existing regime of data protection. Unfortunately, the document does not provide any hint as to what could replace informed consent as the cornerstone of a future regulatory regime for AI. In particular, it is not clear which of two opposing strategies the group recommends:a) improving informed consent procedure, with the goal of ascertaining that online behaviours correspond to authentic acts of consent and that they are adequately informed; b) developing an alternative framework of data governance that downplays the importance of informed consent as a pillar of justification for data-based services.This indication would be highly relevant for both policy and business practice. Endorsing strategy(a) could lead to guidelines and regulations that stress the need to simplify the language used in informed consent procedures, as already prescribed by the GDPR and the GDPR requirements, further. They should provide new criteria for the level of clarity and understanding to be reached, and they should deal with the hard constraint deriving from the fact that people's willingness to spend time managing their privacy is extremely limited. It is thus unclear whether experiments with new ways of conveying information (e.g. short videos?) and of assessing the validity of the process could provide a viable solution (e.g. measurements of the time spent reading privacy policies, short tests to assess their knowledgeability?).Endorsing (b) would lead to reshaping the realm of consumer choices, moving away from assigning a dominant weight to the principle of consumer choice and autonomy in regulatory choices pertaining to consumer privacy. This is, of course, not a new issue. The centrality of informed consent to privacy protection in fair information practices has been the subject of large disputes since these practices have emerged. Critics of informed consent have maintained that informed consent does not substantively limit data collection against the interest of the data controllers, and that it merely provides a perception of privacy protection that is formalistic and enables the accumulation of power by data controllers. This is because citizens often have no choice but to provide their informed consent if they are to access the benefit of certain services. This appears to be still the case in the post-GDPR era, as every person who browses the Internet daily realizes. Those arguing against the centrality of informed consent have always maintained that privacy protection must take a less neutral stance and actively oppose surveillance, beyond attempting to protect efficient exchanges of information. It is not clear if the working group endorses this position or merely asks for better ways to determine if informed consent has been given (leaving it to the research community to solve the problem).5.5. – Potential longer-term concernsThe scientific basis of the assessment of potential long-term harm does not appear anywhere in the document, hence it is difficult to assess the plausibility of this section. On a general note, we believe that the inclusion of merely hypothetical ideas of long-term harm lacking scientific mentioned in the current draft.)On the other hand, relying entirely on experts and whistleblowers is inadequate due to the complexity of some problems arising from the inaccuracy and unfairness of AI. Some of these may be hard to predict ex-ante, before specific real world biased outcomes and predictions arise, solely based on an analysis of the AI system and the data on which it is trained. Some ex-ante assessments are going to be especially problematic in the case of neural networks due to the 'black box' of model explainability. Thus it is also important to promote a sensitivity to AI ethical issues directly in the potentially affected population.Finally, due to the diffusion of neural networks and other 'black box' algorithms, if every AI should comply with the values of traceability and auditability, this could mean that most of current AI development which don't comply with these values is a dead end if the guidelines' principles are taken seriously. Therefore, a common acceptance of these values in public and business sectors is not likely. The guidelines may clarify how this challenge could be tackled. Minor comments:"Human Autonomy" is not a phrase that is normally used in the literature – presumably this is to distinguish it from machine autonomy, but the phrasing is odd.Section (6) on human autonomy includes a prescription about nudging that seems unrealistic and unfeasible, namely, that the functionality of an AI system (which is typically a technology applicable to a limited domain) takes into account the 'overall wellbeing of the user' in the specification of its functionality - e.g. what it nudges the user to do. But measuring the 'overall well-being of a person' is notoriously difficult as a century-old debate in economics testifies. How can this recommendation be helpful to 'realize trustworthy AI' is therefore not really clear.

the question of data resources, it is remarkable that the drafters of this document steer away from mentioning that fair access to data resources is one of the fundamental questions of justice of the data economy. Large data-driven companies, especially US corporations that have in data their largest economic assets, have accumulated a wealth of data on European citizens whose potential for social benefit is underexplored and underexploited. On the one hand, it is difficult to apply fair rates of taxation to these companies, which are able to 'shop around' for the most favourable rates. On the other, the potential of the data to benefit society is limited because the data is stored in their silos. European society could benefit from a bold proposal on how to make the data resources accumulated by these companies to work for the benefit of EU societies. In particular large companies collecting big data about large populations may have information that, if made more broadly accessible, could be used to develop AI-driven innovation in the public sector, including in the context of scientific research.

evidence risks to weaken the overall respectability and practical relevance of this document.

Anonymous	Anonymous	Anonymous	SummaryOur main concern goes to the circumstances and degree of commitment of a non-binding implementation of the Ethics Guidelines for Trustworthy AI as non-obligatory prerequisite of developing and implementing software with a tremendous impact, especially for persons concerned and often labelled as especially vulnerable and exposed to a high level of asymmetry of powers. What to do with applications which logics and flow procedures are not publicly available and was generated without any inclusion or participation of the concerned persons at any stage of the design, the development, the testing and implementation? Our answer or better, requirement: go back to the very start and render the "Ethics Guidelines Trustworthy AI" generally strictly obligatory (i.e. non-voluntary) for any (re-)approach.			The storyOn October 18th, last year epicenter.works (https://epicenter.works), an Austrian NGO in the field of digital-rights, launched (https://fragdenstaat.at/anfrage/quelltext-der-bewertungssoftware-die-im-ams-zum-einsatz-kommt-um-die-perspektiven-aller-arbeitslosen-in-osterreich-zu-bewerten/) a request for information at the Federal Ministry of Labour, Social Affairs, Health and Consumer Protection regarding news (https://futurezone.at/netzpolitik/der-ams-algorithmus-ist-ein-paradebeispiel-fuer-diskriminierung/400147421) reporting (https://derstandard.at/2000089095393/AMS-bewertet-Arbeitslose-kuenftig-per-Algorithmus) "AMS to use algorithm to evaluate unemployed persons in future". The only result was the publication of a paper called "Konzeptunterlage 'Das AMS-Arbeitsmarktchancen-Modell'" (http://www.forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_%20dokumentation.pdf), although the request also asked for information on an audit of the software covering the issues of data-integrity checks, accuracy, long-term-effects, etc. (and ultimately the source-code) - i.e. very similar questions to those raised in the preliminary Draft-Ethics-Guidelines' Assessment List.Although we are not the Assessment List's primary target audience (pdf-page 24) we take the opportunity to check the proposed list addressing the requirements for Trustworthy AI with our knowledge so far on "AMS-Algorithmus". Due to the tremendous lack of (public) information (see above) the replies/answers also suffer from these information deficiencies ("Info not available."- count: 59).The emergence of the "AMS-Algorithmus" is a show-case for an instance of a "worst-case-scenario" in case of non-binding, voluntary Ethics Guidelines.	CheckThe "Assessment List" attached to the Draft Ethics Guidelines was used against the recently introduced "AMS-Algorithmus", an automated system used by the Austrian "Public Employment Service Austria (AMS)" for assessing and evaluating chances of unemployed persons to (re-) gain jobs on the labour-market. The answers are written at the end of the questions.1. Accountability:- Who is accountable if things go wrong? Supposedly the ultimate accountability lies with the AMS entities (e.g. board of directors, supervisory board, etc.) and the Federal Ministry of Labour, Social Affairs, Health and Consumer Protection. - Are the skills and knowledge present in order to take on the responsibility? (Responsible AI training? Ethical oath?) Info not available.- Can third parties or employees report potential vulnerabilities, risks or biases, and what processes are in place to handle these issues and reports? Do they have a single contact point to turn to? Info not available.- Is an (external) auditing of the AI system foreseen? Info not available.- Was a diversity and inclusiveness policy considered in relation to recruitment and retention of staff working on AI to ensure diversity of background? Info not available.- Has an Ethical AI review board been established? A mechanism to discuss grey areas? An internal or external panel of experts? Info not available.2. Data governance:- Is proper governance of data and process ensured? What process and procedures were followed to ensure proper data governance? Info not available.- Is an oversight mechanism put in place? Who is ultimately responsible? Info not available.- What data governance regulation and legislation are applicable to the AI system? Info not available.3. Design for all:- Is the system equitable in use? Info not available.- Does the system accommodate a wide range of individual preferences and abilities? Info not available.– Is the system usable by those with special needs or disabilities, and how was this designed into the system and how is it verified? Info not available.- What definition(s) of fairness is (are) applicable in the context of the system being developed and/or deployed? Info not available.- For each measure of fairness applicable, how is it measured and assured? Info not available.4. Governing AI autonomy:- Is a process foreseen to allow human control, if needed, in each stage? Info not available.- Is a "stop button" foreseen in case of self-learning AI approaches? In case of prescriptive (autonomous decision making) AI approaches? Info not available.- In what ways might the AI system be regarded as autonomous in the sense that it does not rely on human oversight or control? Info not available.- What measures have been taken to ensure that an AI system always makes decisions that are under the overall responsibility of human beings? Info not available.- What measures are taken to audit and remedy issues related to governing AI autonomy? Info not available.- Within the organisation who is responsible for verifying that AI systems can and will be used in a manner in which they are properly governed and under the ultimate responsibility of human beings? Info not available.5. Non-discrimination:- What are the sources of decision variability that occur in same execution conditions? Does such variability affect fundamental rights or ethical	SummaryOur main concern goes to the circumstances and degree of commitment of a non-binding implementation of the Ethics Guidelines for Trustworthy AI as non-obligatory prerequisite of developing and implementing software with a tremendous impact, especially for persons concerned and often labelled as especially vulnerable and exposed to a high level of asymmetry of powers. What to do with applications which logics and flow procedures are not publicly available and was generated without any inclusion or participation of the concerned persons at any stage of the design, the development, the testing and implementation? Our answer or better, requirement: go back to the very start and render the "Ethics Guidelines Trustworthy AI" generally strictly obligatory (i.e. non-voluntary) for any (re-)approach.

principals? How is it measured? Info not available. exactly, but there are some indications in "Konzeptunterlage 'Das AMS-Arbeitsmarktchancen-Modell' for discrimination by gender, etc.- Is there a clear basis for trade-offs between conflicting forms of discrimination, if relevant? Info not available.- Is a strategy in place to avoid creating or reinforcing bias in data and in algorithms? Info not available. exactly, but listening to Mr Kopf, one of the directors of AMS, there seems to be not even comprehension of the very notion of "bias"at the AMS, let alone the idea of how to avoid of creating or reinforcing it.- Are processes in place to continuously test for such biases during development and usage of the system?Info not available. - Is it clear, and is it clearly communicated, to whom or to what group issues related to discrimination can be raised, especially when these are raised by users of, or others affected by, the AI system? Info not available.6. Respect for Privacy:- If applicable, is the system GDPR compliant? Info not available.- Is the personal data information flow in the system under control and compliant with existing privacy protection laws? Info not available.- How can users seek information about valid consent and how can such consent be revoked? Info not available.- Is it clear, and is it clearly communicated, to whom or to what group issues related to privacy violation can be raised, especially when these are raised by users of, or others affected by, the AI system? Info not available.7. Respect for (& Enhancement of) Human Autonomy:- Is the user informed in case of risks on human mental integrity (nudging) by the product? Info not available.- Is useful and necessary information provided to the user of the service/product to enable the latter to take a decision in full self-determination? Info not available. exactly, but according to a letter sent by Viennese administration to the Federal Ministry of Labour, Social Affairs, Health and Consumer Protection in November 2018 there didn't exist any information for users at all.- Does the AI system indicate to users that a decision, content, advice, or outcome, is the result of an algorithmic decision of any kind? Info not available.- Do users have the facility to interrogate algorithmic decisions in order to fully understand their purpose, provenance, the data relied on, etc.? Info not available.8. Robustness:Resilience to Attack:- What are the forms of attack to which the AI system is vulnerable? Which of these forms of attack can be mitigated against? Info not available.- What systems are in place to ensure data security and integrity? Info not available.Reliability & Reproducibility:- Is a strategy in place to monitor and test that my products or services meet goals, purposes and intended applications? Info not available.- Are the used algorithms tested with regards to their reproducibility? Are reproducibility conditions under control? In which specific and sensitive contexts is it necessary to use a different approach? Info not available.- For each aspect of reliability and reproducibility that should be considered, how is it measured and assured? Info not available.- Are processes for the testing and verification of the reliability of AI systems clearly documented and operationalised to those tasked with developing and testing an AI system? Info

not available.- What mechanisms can be used to assure users of the reliability of an AI system? Info not available.Accuracy through data usage and control:- What definition(s) of accuracy is (are) applicable in the context of the system being developed and/or deployed? Info not available.- For each form of accuracy to be considered how is it measured and assured? Info not available.- Is the data comprehensive enough to complete the task in hand? Is the most recent data used (not out-dated)? Info not available.- What other data sources / models can be added to increase accuracy? Info not available.- What other data sources / models can be used to eliminate bias? Info not available.- What strategy was put in place to measure inclusiveness of the data? Is the data representative enough of the case to be solved? Info not available.Fall-back plan:- What would be the impact of the AI system failing by: Providing wrong results? Being unavailable? Providing societally unacceptable results (e.g. bias)? Info not available.- In case of unacceptable impact - Have thresholds and governance for the above scenarios been defined to trigger alternative/fall-back plans? Info not available.- Have fall-back plans been defined and tested? Info not available.9. Safety:- What definition(s) of safety is (are) applicable in the context of the system being developed and/or deployed? Info not available.- For each form of safety to be considered how is it measured and assured? Info not available.- Have the potential safety risks of (other) foreseeable uses of the technology, including accidental or malicious misuse thereof, been identified? Info not available.- Is information provided in case of a risk for human physical integrity? Info not available.- Is a process in place to classify and assess potential risks associated with use of the product or service? Info not available.- Has a plan been established to mitigate and/or manage the identified risks? Info not available.10. Transparency:Purpose:- Is it clear who or what may benefit from the product/service? From our point of view: No.- Have the usage scenarios for the product been specified and clearly communicated? From our point of view: No.- Have the limitations of the product been specified to its users? Info not available.- Have criteria for deployment for the product been set and made available to the user? Info not available.Traceability:- What measures are put in place to inform on the product's accuracy? On the reasons/criteria behind outcomes of the product? Info not available.- Is the nature of the product or technology, and the potential risks or perceived risks (e.g. around biases) thereof, communicated in a way that the intended users, third parties and the general public can access and understand? From our point of view: No.- Is a traceability mechanism in place to make my AI system auditable, particularly in critical situations? Info not available.This entails documentation of:o Method of building the algorithmic system- In case of a rule-based AI system, the method of programming the AI system should be clarified(i.e. how they build their model) From our point of view: No.- In case of a learning-based AI system, the method of training the algorithm should be clarified. This requires information on the data used for this purpose, including: how the data used was gathered; how the data used was

selected (for example if any inclusion or exclusion criteria applied); and was personal data used as an input to train the algorithm? Please specify what types of personal data were used. Info not available.o Method of testing the algorithmic system- In case of a rule-based AI system, the scenario-selection or test cases used in order to test and validate their system should be provided. Info not available.- In case of a learning based model, information about the data used to test the system should be provided, including: how the data used was gathered; how the data used was selected; and was personal data used as an input to train the algorithm? Please specify what types of personal data were used. Info not available.o Outcomes of the algorithmic system- The outcome(s) of or decision(s) taken by the algorithm should be provided, as well as potential other decisions that would result from different cases (e.g. for other subgroups). Info not available.

Marta Rocchi, Pierangelo Rosati, Theo Lynn	M.Rocchi, P. Rosati, T. Lynn	Irish Institute of Digital Business - Dublin City University Business School	p. 1. We endorse the importance of AI being "human-centric" instead of "human-centred." The emphasis posed in the Draft on the "human-centric" dimension of AI as essential is antecedent to the fact of being "trustworthy." More consideration, however, needs to be given to how trust will be built between humans and AI and in particular addressing concerns relating to benevolence, integrity, competence and predictability.p. 2 "The Role of Ethics"The document states that: "Ethics as a field of study is centuries old and centres on questions like 'what is a good' action, 'what is right', and in some instances 'what is the good life'". Greater emphasis needs to be given to ethics as a field of inquiry aimed at understanding "what is the good life" in the context of the impact of the implementation of AI for society as a whole. This would reflect a more agent-centric approach, as opposed to an act-centred approach to ethics.p. 3: We endorse the idea that AI is context-specific, so there is a need for a tailored approach for its use in different sectors.Further consideration needs to be provided in the context of a multi-stakeholder (citizens, regulators and AI providers) approach to assurance and accountability of AI with appropriate declarative, confirmative, preventative, detective, and corrective controls.	Point 1. "The EU's Rights' Based Approach to AI Ethics" The notion of "common good" is worthy of more attention in so far as one of the risks of AI-based systems is that they may only be beneficial for one segment of the population, not for each and every one. If the fourth industrial revolution is to avoid the negative outcomes of the first and second industrial revolution, rediscovering and placing greater emphasis on the concept of the common good is needed. There should be equality between those in society who benefit from the implementation of AI-based systems and those who pay the cost in terms of quality of labour and quality of life at both a macro and micro-level. For example, at a micro- and nano-level, more consideration is needed with regards to the "human cost" of training AI (psychological effects coming from the exposure to long hours of videos, extreme or not desirable images, etc). This is consistent with Sustainable Development Goal n.8 and Sustainable Development Goal n.10 relating to "Decent Work and Economic Growth" and "Reduced Inequalities."Point 2. From Fundamental Rights to Principles and ValuesThere is an almost absolute focus on the "principle of autonomy." It would be interesting to complement and integrate the "principle of autonomy" with the concepts of "dependence," "vulnerability," and "relationality." Autonomy is not the specific characteristic of human beings: thinking about a baby or an elderly person impeded in her movements or in her way of reasoning, we cannot deny that these people are fully human beings, even if they are not autonomous in the sense described in the document (i.e. "free to make choices about their own lives, be it about their physical, emotional or mental wellbeing," p. 5). The document makes reference to the principle of autonomy as descendant from the fact that human beings are free. Rather, all human beings are free but not all human beings are necessarily autonomous. A "human-centric" AI should take into account a principle that is as generalizable as human dignity is. We suggest the consideration of "relationality" or "dependence"/"vulnerability" to be integrated or to complement the "principle of autonomy." Given that the document makes explicit reference both to the pursuit of the good life and the common good, highlighting a relational component of human beings would be more in line with those concepts, instead of relying merely on autonomy.Point 3: Fundamental Rights of Human Beings3.4 A massive introduction of AI-based systems can create an unequal social context. So equality, in the context of the pursuit of a common good, means that costs and benefits are equally distributed and accessible.3.5 The text reports the following: "Citizens should never be subject to systemic scoring by government". Later in the text, section 5.3 reports that scoring has been often used for example in schools, e-learning, or driving licences. In light of the discrepancy between what is stated in section 5.3 about the practice of scoring, and what is stated in point 3.5 about how scoring should (not) be, a clarification is needed. What does it mean "systematic" in this context? Additionally, governments cannot engage in systematic scoring, so who can actually do it? The difference seems to be the extent of the domain for scoring,	1. Requirements of Trustworthy AI1. Accountability: There is a mismatch between the title and the content under the point on "accountability." Further consideration needs to be provided in the context of a multi-stakeholder (citizens, regulators and AI providers) approach to both assurance and accountability of AI with appropriate declarative, confirmative, preventative, detective, and corrective controls.2. Data governance: this paragraph seems to focus on basic best-practices for analytics rather than on data governance. It does not seem to cover (even if at general level) the entire data lifecycle. We suggest to integrate this section with explicit references to security, quality monitoring, data deletion, etc.3. Design for all: this concept is expressed in a very ambitious way. We do not deny the desirability of this outcome, and again the use of the expression "user-centric" instead of "user-centred" helps capturing the idea that the real source of information is each specific person, not an ideal person, who is not really close to anybody in particular, unless to the "average" person. Is it really possible to always design AI-based systems that are accessible for everyone? Moreover, this paragraph seems to communicate the idea that every citizen should be able to access/use complex systems, but in reality they will either provide inputs to or consume the output of these systems. Accessibility is context-related as it depends on who is going to be target user of a system.4. Governance of AI autonomy (human oversight) – See above. It might be worth making more explicit what controls will be put in place.8. Robustness > Reliability and reproducibility. In relation to: "the accuracy of results can be confirmed and reproduced by independent evaluation," is this practically doable? How could anyone replicate the results of a system being trained for years (at a reasonable cost)?8. Robustness > Resilience to attack. This wording could result to be confusing. "Security" would be more accurate as systems should be resilient to (human and, in the context of AI also, non-human) errors; not only to "attacks."10. Transparency The title suggests an objective which is probably too ambitious. We wonder whether full transparency in the commercial context is doable/desirable. The first sentence of this section seems to take this into account when referring to "reduction of information asymmetry." This point partially overlaps with reliability/reproducibility.2. Technical and Non-Technical Methods to achieve Trustworthy AIAccountability GovernanceAgain, a framework is needed for both assurance and accountability that comprises both technical and non-technical methods. Accountability is useful for assigning blame and corrective actions. Trust mechanisms need to be designed with AI in mind, including feedback reputation systems, third party assurances, AI transparency mechanisms, AI performance verification systems, 'formal' trust mechanisms and even trust labels for communicating the trustworthiness of AI attributes of a given AI system.We support the idea that technical and non-technical methods are required. Controls should address each aspect under consideration: ethics, rights, regulation etc. and who the appropriate stakeholders for each aspect should be.	-	The definition of AI provided in the glossary seems general enough to include a wide range of Machine Learning/AI applications. However, given that this is the main focus of the document, a more extended and detailed definition could have been presented in the main text. Adding explicit reference to some uses/cases of current and future/potential implementation could be useful. Exemplar uses/cases could focus on "extreme" applications to better clarify the range of applications the guidelines aim to cover.Also in the parallel HLEG document on the definition of AI (AI HLEG "A Definition of AI", 2018), we found spaces for improvement in the characterization of intelligence as a "vague concept" (p. 1). The dialogue on Artificial Intelligence can open a space to better understand and characterize human intelligence, with a grounded anthropological basis. An anthropology based on relationality and positive dependence (vulnerability) instead of or as an integration to autonomy could help in defining what is human intelligence.Wording: the document often refers to "consequences" of A.I. Using the term "impacts" may better express what the document is aimed at covering, as it can include both 'direct' and 'indirect' effects. "Consequences" could imply a direct causal relationship.

however a clarification note would be helpful. For example, implementing domain-specific scoring (even at a small scale) is possible, but it could still be not desirable if done for not a good purpose.As per previous comments, a multi-stakeholder framework for assurance and accountability of fundamental rights is necessary.Point 4: Ethical Principles in the Context of AI and Correlating ValuesOn the Principle of Beneficence. - P. 8. The document states “AI systems should be designed and developed to improve individual and collective wellbeing” – we would keep coherence with the first part of the document, keep mentioning the common good. Wellbeing is part of the common good, but it is not sufficient to describe it.- P. 9. The document says “by helping to increase citizen’s mental autonomy.” As already stated, we suggest using another kind of framework instead of that of the autonomy. One of the challenges of living in a AI-based society is to keep one’s own critical perspective on reality and ability to choose. Instead of the “mental autonomy,” it can be rather helpful to state something like: “by helping to increase citizen’s critical thinking.”It could be helpful to give greater consideration of benevolence vs beneficence and whether they ethically and/or psychologically represent the same concept.On the Principle of Autonomy.As already stated, we suggest keeping the “preserve human agency,” instead of declaring autonomy as a principle. Alternatively, the “principle of relational autonomy” or “integrated autonomy” can be examined.On the Principle of Explicability: “Operate Transparently”This section states that “business model transparency means that human beings are knowingly informed of the intention of developers and technology implementers of AI systems.” Again, further consideration needs to be provided in the context of a multi-stakeholder (citizens, regulators and AI providers) approach to assurance and accountability of AI with appropriate declarative, confirmative, preventative, detective, and corrective controls. Such controls would ensure explicability and transparency.Point 5: Critical concerns raised by AI5.1 Identification without consentCitizens should not be identifiable without consent unless for justifiable safety or legal reasons.5.2 Covert AI systemsCentral to trust is transparency. Again, this goes to the core of assurance and accountability of AI. What are appropriate declarative, confirmative, preventative, detective, and corrective controls to ensure that covert AI systems are not used for unethical purposes? In light of this section, we would recommend to consider our proposal of calling this a “human-centric and trustworthy AI” instead of just a “trustworthy AI.” This will help re-focus the attention on the primacy of humans over androids and humanoids. For example, in a scenario of limited resources, there should be given priority in healing a human being rather than an android.

Merel	Noorman	Tilburg Institute for Law, Technology and Society, Tilburg Law School, Tilburg University	Comments on behalf of the TILT AI and Robotics group:The HLEG has taken on the ambitious project of developing general guidelines for ethical AI and has, a first step in this process, published a draft document for stakeholders to review and comment upon. In this draft document, the group builds upon, and rightly so, a range of existing frameworks, principles and manifestos. It proposes to centre the guidelines on the concept of Trustworthy AI. They elaborate this concept in three sections that each address different levels of abstraction: ethical purpose rooted in fundamental rights, technical and non-technical methods, and an assessment list. We would like to congratulate the HLEG on this first step in a complex and multifaceted process and complement the group on finding a shared basis to further build upon. In particular, we welcome the rights-based approach that HLEG chose to pursue, as it roots the guidelines in shared values and principles within Europe while at the same time aligning them with many of the existing guidelines. Moreover, we were pleased to see the substantive definition of AI as it is outlined in the document published in parallel with the guidelines and summarized in the draft document. In particular, by distinguishing between AI as a technology and artefact designed and deployed by human beings on the one hand and AI as a scientific discipline on the other, the authors have managed to highlight the extensiveness and heterogeneity of AI. They have also signalled the human agency and work that is involved in making these AI systems function. The focus in the definition on the pre-determined goals and parameters provides regulators something to work with. The HLEG also brings the discussions on ethical AI a step forward by not only focusing on rights, principles and values, but also on the implementation and embedding of the technology. The ambition to provide concrete tools and methods for policy makers, developers, and citizens is needed to bring ethical AI into practice and we encourage further work in this direction. As the HLEG has explicitly asked for critical feedback we would like to offer a few suggestions and comments for the further improvement of the document. We will first provide some general comments and then go into more specific comments per section of the guidelines document. General (conceptual) comments Our general comments focus primarily on the conceptual elaboration of key terms and ideas in the document. In particular, we argue that the concept of trustworthiness needs to be centred on vulnerability and uncertainty; AI should be treated as embedded in a larger sociotechnical context and the benefits and risks of AI require a more nuanced consideration. TrustworthinessThe HLEG has chosen to centre the guidelines on the concept of trustworthiness, which has the potential to provide a useful tool for the different audiences of the document to structure their thinking about how to proceed (or not) with AI. Trust, according to the HLEG, is the cement of societies, communities, economies and sustainable development. Early on in the document, the HLEG defines Trustworthy AI as consisting of two components: "(1) its development, deployment and use should respect fundamental rights and applicable	While applauding the rights-based approach of the HLEG, we would encourage the authors to avoid the conflation of fundamental rights and ethics in the concept of 'ethical purpose'. Law and ethics are two separate domains that need to be clearly distinguished with regard to their rationale and function. Not doing so, runs the risks of obscuring or down-playing the central role of law in the governance of the design, deployment and use of AI and to reinterpret ethics as 'industry self-regulation'. Yet, ethics should go above and beyond the law. Moreover, we would like to note that the authors spend some time on explaining the cycle from rights to principles to values. However, they let this cycle go in the following chapters.A minor point to be made here, is that on page five the authors suggest that they derive the principle of autonomy from human dignity, but it seems to us that this principle should derive from freedom and liberty. Our further comments on this section focus in particular on the interpretation of the fundamental rights, the selection and description of the principles and the reasoning behind the critical concerns on AI.Fundamental rightsThe Chapter on ethical purpose that elaborates on the fundamental rights requires re-thinking and re-structuring. In particular, the section on fundamental rights lacks a firm logical structure. It is not clear what the authors are basing their decisions on for their interpretations of these rights. We recommend that the HLEG links the description to existing frameworks in a more systematic and objective way.The description of the fundamental rights currently appears to be an ad hoc creation. In particular, paragraph 3.4 seems peculiar and does not match with existing legal frameworks. The five sentences in this section contradict each other. Equality does not mean equal treatment of everyone regardless of the situation. Rather, it means people should be treated equal in equal situations and unequal in unequal situations according to their unequalness. Equality in AI should be about neutrality in access and how it applies to you and affects you. Combining equality and the rights of minorities, might make sense from a natural language perspective, but not necessarily from a legal perspective: if all situations are treated equally, it is impossible to protect minorities, which by definitions are in a different position than the majority and require a different treatment. Another curious element in this section is the suggestion that consumers and workers are minorities. We assume that the authors intended to note that equality requires respect for the position of less powerful stakeholders, such as consumers and workers. At the same time, it is a bit odd to find the word consumer in a human rights framework instead of citizen. Section 3.5 about citizen rights is suddenly very specifically aimed at the public sector, while corporations are completely left out. In our view, this is an omission. Citizens must be protected from preventive intervention by corporations. AI should not be employed to inhibit citizens' rights in their relations with public and commercial institutions alike.What is also missing in this section - in fact it is mostly missing from the entire document - is the respect for human relations and the environment. The fundamental rights section	Chapter 2 provides an overview of possible methods to implement the ten requirements that have been derived from the principles in the previous Chapter, including accountability, data governance and respect for privacy. Although these requirements are certainly some that AI developers and users should adhere to, it is difficult to evaluate these principles at the given level of abstraction. Moreover, it is again unclear why the HLEG chose these particular requirements and left others out. For example, why is environmental sustainability not part of this list? Also, as part of the principle of explicability the authors mention that informed consent should be a requirement, yet it is absent from the list and does not come back as part of the requirement of transparency. Similarly, robustness is a requirement and thus elaborated, but reliability is not. With regard to the requirement of accountability, it should be noted that accountability is not just about compensation, but also about learning and adjusting existing practices in order to prevent or minimize the risk of untoward events from occurring again. Here again, it would be helpful if the authors could provide some reasons for choosing particular interpretations of concepts. In the section on non-discrimination, they provide some references in support of particular definitions. They might want to do this for the other sections as well. The relations and possible conflicts between the requirements also warrant further elaboration. For example, the authors illustrate the relation between other values such as non-discrimination. However, they do not mention the potential conflicts between privacy and identifying and correcting problematic bias. Nor do they address the potential conflict with transparency or the supplementary, mutual support between privacy and human autonomy or safety. To implement the requirements, the authors provide a list of technical as well as non-technical methods. Although potentially helpful as a starting point, we offer a few suggestions for consideration: - Although audibility is mentioned in the technical section, it is missing from the non-technical section. No mention is made of the kind of institutions or mechanisms that are necessary to audit these technologies. Moreover, the authors may want to further elaborate the notion of auditability.- Democratic decision-making is missing from both the technical and non-technical methods.- The support of interpersonal relationships to foster trust is missing. - Learning and training with new systems is missing. - Developing new protocols for the deployment and use of AI etc. could be added to the list of non-technical methods.	In the last part of the document the HLEG provides an assessment list "to operationalise the implementation and assessment of the requirements of Trustworthy AI set out above, throughout the different stages of AI development and use." This is a potentially helpful way of providing a concrete tool for the intended audience to work with and there is definitely a demand for such a list. However, we feel that such a list or set of lists would be very context sensitive and it is therefore difficult to comment on this rather abstract list without the necessary context. Nevertheless, we would like to point out a few issues that may inform future work on the assessment list(s). In particular, one question that is not in the assessment list is whether the use of AI is justifiable given the circumstances. Are there other better ways of solving a particular problem? In addition, the document does not discuss (unintended/unanticipated) interactions with other systems nor the embedding of the system in existing practices. For the successful adoption of AI systems, these are things that need to be taken into consideration. Finally, the assessment list is currently lacking an operationalization of the value-sensitive design approach (e.g. stakeholder inclusion, weighing of different values).We would like to conclude by once again congratulating the HLEG on this first step in developing the ethical guidelines. We hope our feedback and suggestions will contribute to further fine-tuning of the guidelines document.	See comments under "Introduction: Rationale and Foresight of the Guidelines".

regulation, as well as core principles and values, ensuring an “ethical purpose”, and (2) it should be technically robust and reliable.” (p. 1) However, the current use of the concept in the document runs the risks of providing another rhetorical tool for parties involved to carry on with business as usual, while claiming to have adopted an ethical approach to AI. The elaboration of trustworthy AI in the draft document suggests there is technological fix for a possible lack of trust, namely ensuring ethical purpose and technological robustness and reliability. Once these two criteria have been met, citizens and others will be able to maximize control and minimize risks and thus not have to worry about AI. Yet, trust is fundamentally about vulnerability and uncertainty. We trust someone when we know there are uncertainties and potential risks, but we are nevertheless willing to work with or rely on them. Unfortunately, this uncertainty and vulnerability is underexposed in the current guidelines and should, in our opinion, be put centre stage. Trustworthy AI should be about how we deal with the uncertainty and vulnerability that come with the development, deployment and use of AI. Technological reliability and robustness are to be encouraged, but what happens when things go wrong? What gives us that trust that things will work out despite the uncertainties and vulnerabilities? Are citizens sufficiently informed about the potential risks of AI technologies? Is it only the AI system that should be reliable, or should the sociotechnical system in which it is embedded also be reliable and robust? Will AI systems afford trust in institutions? When should we cultivate a healthy distrust? The guidelines should emphasise that trust is a means of dealing with the unknown.AI as embedded in a larger sociotechnical contextIn part, the narrow conceptualization of trustworthiness, as reflected throughout the document, is the result of a tendency of the authors to treat AI as a monolithic autonomous thing, isolated from its context. On several occasions the authors attribute agency to AI in such a way that is obscures the work done by human beings. In the Executive Summary, the authors note that human beings will only be able to confidently and fully reap the benefits of AI if they trust the technology. However, it is not the technology that we need to trust, as the authors justly note in a later section of the document, when they argue:“Trust in AI includes: trust in the technology, through the way it is built and used by human beings; trust in the rules, laws and norms that govern AI […] or trust in the business and public governance models of AI services, products and manufacturers.” [emphasis added] (p. 2) In our view, the question then is how do (1) the human beings that build and use these technologies; (2) the rules, laws, and norms that govern these activities, and (3) the business and public governance models for these technologies deal with uncertainty and vulnerability in such a way that they foster trust? Although focusing on ethical purpose and technical robustness and reliability, are part of this, it is not sufficient. Ensuring that the design, development and deployment of AI respect rights and regulation, and adhere to core principles and values, and making technology robust and reliable, will not simply dissolve the uncertainties and vulnerabilities. Although, is a key part of the guidelines document. The authors note that “[c]ompliance with these Guidelines in no way replaces compliance with [fundamental rights and with all applicable regulations], but merely offers a complement thereto" (p.2). It therefore needs to provide a solid foundation and should be linked explicitly to existing fundamental rights. Particular interpretations should not depart from those that are accepted within these frameworks. Yet, in its current form, the document seems to argue for a lower standard then is currently set by the existing legal framework.Ethical principlesThe HLEG proposes five principles for AI: beneficence, non-maleficence, autonomy, justice and explicability. The first four principles are well known leading principles in the medical, care and bioethics domains. As such, these principles are in line with the principles put forward by existing ethical principles in various fields, including computer ethics, as well as principles offered by various proposed AI ethical guidelines. The HLEG has added the fifth principle of explicability. These principles are intended to guide the operationalisation of core values derived from the fundamental rights.It is curious that AI would turn to the four principles of Beauchamp and Childress given the rigorous controversy that surrounds both the usefulness and insufficiency of “The Four Principles” as a moral framework. In 1995, leading UK bioethicist, Soren Holm argued that “The theory [the four principles of bioethics] is developed as a common-morality theory, and the present paper attempts to show how this approach, starting from American common-morality, leads to an underdevelopment of beneficence and justice, and that the methods offered for specification and balancing of principles are inadequate.” The reference to these principles without incorporation of the fruits of the robust critique that has ensued in the past twenty years and has led to more nuanced and useful engagement with ethical principles and what they require is a missed opportunity. Greater nuance and attention to the widely-recognised limitations of the four principles is strongly encouraged.The authors have chosen to highlight the principle of explicability for AI, which is understandable given the seeming complexity of AI systems. The authors define explicability in terms of transparency, where transparency is about the auditability, comprehensibility and intelligibility of AI systems as well as about the awareness of the intentions underlying particular business models. This kind of transparency, according to the HLEG, is needed for informed consent. Explicability, according to the document, is primarily about consent as a form of control for citizens. Yet, on its own consent is a very weak instrument, and should not form the primary basis for requiring these practices.The strong connotation with the medical and bioethics domain also makes us doubt whether the addition of the fifth principle of 'explicability' is the right choice. The notions of explaining (and understanding) are very much already a part of the 'mother' framework, notably, the autonomy of patients/subjects is served through the informed consent paradigm. Where the other principles are 'ends in themselves', explicability is not. The side effect of singling out explicability is that it

the authors address uncertainties and vulnerabilities in Chapter 2 - for instance through the requirement of accountability and data governance - the two elements should be at the heart of the definition of trustworthy AI. The author's treatment of AI is also problematic in the operationalization of trustworthy AI. Despite the nuanced definition that it has been given in the separate document, the draft document does not position AI - as a system or scientific discipline - within the broader discussion on the ethics of digital and data-driven technologies. AI techniques are rarely developed or used as standalone systems, but are embedded in broader eco-systems. They are components of decision-making processes that extend throughout and beyond organizations and involve multiple human beings in different roles. AI techniques, are for example, used by platform companies to analyse data obtained through social media to optimize advertisement targeting for different companies. Many critical analyses have been made of the network dynamics involved in these developments, e.g. monopolization, power asymmetries, etc.. Yet, the problem of growing power asymmetries as a result of large scale datafication of society, partly enabled by AI technologies, is not discussed in the document. Or perhaps, one could argue, that it is addressed in the requirement of respect for human autonomy, but only in a very abstract sense without reference to other developments that fuel these asymmetries. By positioning AI as an isolated thing, we lose sight of all the elements in the sociotechnical systems that contribute to its trustworthiness, such as human relationships and stabilizing institutions. Benefits and risksFinally, the authors suggest in the Executive Summary that on the whole the benefits of AI outweigh its risks and we should therefore invest in maximizing the benefits while minimizing the risks. However, given the abstractness of AI as used in the document, this is an empty and misleading statement. It suggests that there is some way of measuring the benefits and risks of this very ambiguous and broad thing called AI and that there is a moral imperative to pursue the benefits. It places the critical citizen, company or organization in the awkward position of not being able to refuse the technology or opt-out of its use. AI has to be used, and therefore we should make it trustworthy. This precludes a lot of other options. Perhaps this is also a reason for the absence of a discussion on the precautionary principle in the document.

diminishes the importance of informed consent in the principle of autonomy. By separating it from the principle of autonomy, it narrows human autonomy to the ability to choose to opt in or out and not much more. Yet, the kind of transparency that the HLEG proposes should empower human beings to not only be able to choose to opt-in or out, but to understand why one might choose to do so, under what conditions and how to object to the framing of the offered choice.Moreover, the principle of explicability does not reflect the current state of AI and our current limited understanding of what explicability should mean. It is key that we will find ways to make these technologies sufficiently understandable to serve responsible use, but as we have not yet defined how to serve this understanding, it is unclear what 'explicability' is or should be. Variants of explanation may thus be offered to comply with this principle, and it will be hard to argue why these will or will not serve the purposes as contained in the other principles. At the same time, the principle of explicability seems to undermine the importance of trust, as trust is needed when things are not transparent; when you do not know the innerworkings of something. If the focus is on trust, the question is how do we deal with not knowing for sure? It is not enough to have an indication that AI will function as expected; trust also comes from knowing that there is a safety net when things do not work out as expected. This is one important reason to have a strong legal framework. If AI and the human beings behind it are able to explain what the systems does and why, then it helps citizens if they can legally hold those responsible to their explanation. If AI raises certain expectations, it is important that citizens can legally hold those people behind the AI accountable for the impression the AI made. The law functions here as a safety net. Citizens are not left to their own devices, if things go wrong and AI and the human beings behind prove not worthy of citizens' trust. In terms of human autonomy, we suggest that the authors reflect on the connection between human and AI autonomy. On the one hand to clearly distinguish between the two, but on the other hand to also signal the importance of human autonomy in meaningful human control (Human oversight of AI autonomy). Although citizens can be attributed responsibility for the behaviour of certain systems, they have to be in the position to exercise their autonomy in order to be able to live up to the responsibility. That is, they should not only be able to understand AI, but they should also have sufficient discretionary power and be able to intentionally affect the behaviour of the system directly or indirectly. Moreover, it should be noted that it is generally not one individual that is responsible. Responsibility should be appropriately and fairly distributed across stakeholders in accordance with the level of control or influence they have. One final note on the principles. Although solidarity is mentioned in passing, it is underdeveloped in the description of the principles. The emphasis is on human individuals, but not on the relationships and common bonds between them. This disregards the influence of AI on the social whole(s), while placing too much burden on the individual as the actor that needs to

know/trust/understand. We look forward to seeing the further elaboration of the case-studies that the HLEG has announced for the following version of the document. This would be a valuable contribution to the document as it will help to demonstrate how certain rights and principles are relevant and applicable to particular AI designs and uses, as the HLEG notes. For example, the rule of law might not be that relevant for a smart-refrigerator, but key to law-enforcement AI systems. Similarly, not every system should have an opt-out option, but those that do not have an opt-out should adhere to much stricter requirements in terms of the rule of law. Moreover, it would allow the authors to address the issues of potential conflicts between principles as well as between rights. With regard to the domains chosen for case-studies, public administration systems seem to be conspicuously missing.Critical concerns raised by AIIn the final section of this Chapter, the HLEG presents a list of possible concerns about AI and notes that this has proven to be a contentious part of the process. Although we appreciate the concerns about the future developments in AI, it is unclear what the connection of this list of somewhat generic scenario’s is to trustworthy AI. Moreover, what are the reasons for choosing these scenario’s and leaving out others? What is missing, for instance, are economic concerns, such as the influence of AI on the labour market or growing (income) inequality? We would have expected a more systematic evaluation of what the possible risks of the proposed trajectory towards trustworthy AI might be.We would propose to start this section by establishing some criteria for elaborating the critical concerns. One such criterion could be the timeframe to look at. Subsequently, we suggest linking the concerns to the work already done in the previous chapters, and examining how developments in AI could threaten trustworthiness or what would happen if fundamental rights and principles are not respected. For instance, what happens when the principle of autonomy is ignored? A more systematic analysis would connect this section better to the preceding sections and would enrich the elaboration of the concept of trustworthiness. Some minor comments on this section are perhaps also worth noting: - There is no such thing as ""anonymous” personal data" (p. 11).- Only one of the legal bases for personal data processing has been discussed, despite the reference to art. 6 GDPR (idem).

Bojana	Bellamy	Centre for Information Policy Leadership	Introduction — Rationale and Foresight of the Guidelines (pages 1 to 3) The Centre for Information Policy Leadership (CIPL) welcomes the opportunity to submit comments to the EU Commission High-Level Expert Group (HLEG) on Artificial Intelligence on its Draft Ethics Guidelines for Trustworthy AI (the ”Ethics Guidelines” “Draft Guidelines” or “Guidelines”).CIPL shares many of the views expressed in the Draft Guidelines and, in particular, that AI brings significant benefits for users and society and that its use should be facilitated. CIPL fully supports the overall objective to maximise the benefits of AI while minimising the risks.In addition, CIPL welcomes the objective of the HLEG to develop Ethics Guidelines which are not just a compilation of values and principles to be respected, but which provide guidance on how to actually implement them in the development and use of AI systems.Although the Guidelines are intended to be voluntary and non-binding, they will nonetheless play a role in shaping the meaning of certain concepts and terms as they relate to AI and data protection which can be influential in framing the context and design of new regulations. As a result, CIPL suggests some changes and adaptations to the content of the Draft Guidelines.CIPL agrees that trust, powered by an ethical approach to AI and technical robustness, is a prerequisite to fostering a climate favourable to AI’s development, deployment and uptake. This is particularly important for individuals’ willingness to share data as AI relies on the availability, use and analysis of data. In certain instances, a significant amount of data may be required and necessary to appropriately and accurately deliver results. Such data can be personal data or non-personal data although fewer and fewer instances exist were data cannot be traced back to individuals. As a result, data protection laws often come into play within the context of AI.CIPL believes that trust in AI can be promoted by demystifying some unfounded concerns surrounding it. The public needs to be further educated on what AI is, how it can be used and how it can benefit society. In particular, CIPL welcomes the recognition by the HLEG that not only private but also public entities and governments can develop, deploy and use AI for the public good. This provides a strong argument for the promotion of AI in civil society. CIPL fully supports the recognition by the HLEG that “no legal vacuum currently exists” as regulation already applies to AI. CIPL cautions, in this context, against regulating AI as a standalone technology where the legal ecosystem already provides for relevant safeguards, including data protection and privacy laws such as the EU General Data Protection Regulation (GDPR). Similarly, AI should not be stigmatised when compared with other technologies that carry similar legal or ethical challenges.CIPL is currently considering the interplay between AI and data protection through its project on “Delivering Sustainable AI Accountability in Practice”. This project aims to provide a detailed understanding of the opportunities presented by AI, its challenges to data protection laws and practical ways to address these issues through best practices and organisational accountability. CIPL’s first report “Artificial Intelligence and Data Protection in Tension” is available here:	Chapter I – Respecting Fundamental Rights, Principles and Values — Ethical Purpose (page 5)CIPL supports the Guidelines’ reliance on the commitment to fundamental rights of the EU Treaties and the Charter of Fundamental Rights. Article 8 of the Charter, in particular, deals with the protection of personal data which is further safeguarded by the General Data Protection Regulation (GDPR). Such instruments should not, however, be interpreted in an overly restrictive manner in the context of AI as further explained below.3. Fundamental Rights of Human Beings (page 7)3.1 Respect for Human Dignity The Draft Guidelines state that “[i]n the context of AI, respect for human dignity entails that all people are treated with respect due to them as individuals, rather than merely as data subjects”.CIPL recommends revising the current wording which reflects a dogmatic distinction between individuals and data subjects in terms of providing respect and dignity. The notion of “data subject” does not have the pejorative connotation implied by the Guidelines (individuals are treated “merely as data subjects”). Personal data collection and processing is necessary in many aspects of individuals’ everyday life to provide access to a growing number of services with huge individual and societal benefits. “Data subject” is just a term of art used by most data protection laws that is designed to uphold the rights and freedoms of individuals when their personal data is being processed.3.2 Freedom of the IndividualThe Draft Guidelines explain that, in the AI context, this right requires “protection from direct or indirect coercion, surveillance, deception or manipulation”. The Guidelines should qualify that this right is not absolute and in some cases requirements regarding national security obligations of the government to protect individuals in society outweigh such a right. 3.3 Respect for Democracy, Justice and the Rule of LawCIPL supports this principle and its definition — in particular, the need for AI to provide for a right of review, scrutiny and redress for individuals, as such rights will, in some instances, compensate for the possible lack of transparency surrounding specific AI systems.3.5 Citizens RightsThe Draft Guidelines note that “citizens should enjoy a right to be informed of any automated treatment of their data by government bodies, and systematically be offered to express opt-out. Citizens should never be subject to systematic scoring by government”.CIPL cautions against such a broad right for citizens to be informed of the automated treatment of their data and to exercise opt-out as this may run counter to the relevance of the data processing itself (for instance, for tax processing purposes or public security). In this respect, CIPL recommends referring to the standards of the GDPR that also apply to the data processing of public bodies and, in particular, its provisions on automated decision-making (Article 22) (See also further comments below on these topics). Similarly, the point that “[c]itizens should never be subject to systematic scoring by government” should be more nuanced, taking into account specific data processing that are necessary for the public good and provision of services to society. CIPL recommends instead including a right to information and to obtain redress, where	Chapter II – Realising Trustworthy AI (page 14)1. Requirements of Trustworthy AI (page 14)The Draft Guidelines set out ten main requirements for AI to be trustworthy. CIPL supports these requirements and offers the following comments on the requirements of accountability, design for all, non-discrimination, respect for privacy, robustness and transparency:1. Accountability (page 14)CIPL believes that the accountability requirement should be considered as the overarching principle for trustworthy AI. All other principles and requirements should be grounded in the accountability principle. This may already be implied by the HLEG by including accountability as the first of the ten requirements in the Draft Guidelines.The definition of accountability provided on page 14 of the Draft Guidelines appears to be limited to redress mechanisms only. The definition of accountability, as employed by the GDPR and other privacy frameworks, is much broader than merely providing redress mechanisms to individuals, as described below, and CIPL recommends that the Guidelines reflect this. CIPL has worked extensively on the topic of accountability and has published several papers on the central role of organisational accountability in data protection. To read more about CIPL’s work in this area please see the following papers: “Introduction: The Central Role of Organisational Accountability in Data Protection”, available at https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/introduction_to_two_new_cipl_papers_on_the_central_role_of_organisational_accountability_in_data_protection.pdf; “The Case for Accountability: How it Enables Effective Data Protection and Trust in the Digital Society”, available at https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_accountability_paper_1_-_the_case_for_accountability_-_how_it_enables_effective_data_protection_and_trust_in_the_digital_society.pdf and “Incentivising Accountability: How Data Protection Authorities and Law Makers Can Encourage Accountability”, available at https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_accountability_paper_2_-_incentivising_accountability_-_how_data_protection_authorities_and_law_makers_can_encourage_accountability.pdf.As further explained in these papers, accountability essentially requires two things: (1) a comprehensive data protection compliance program implementing all applicable requirements, including those relating to oversight, risk-assessment, policies and procedures, transparency, training, redress, etc. and (2) the ability of an organisation to demonstrate this program to regulators upon request. Articles 5 and 24 of the GDPR and the Article 29 Working Party’s 2010 guidelines on accountability express the principle as requiring organisations to take the necessary steps to implement applicable data protection principles or goals and to be able to demonstrate such implementation. In order to achieve this, organisations will have to implement effective comprehensive privacy management programs that will take into account and address any risks as well as fairness or ethical issues relating to AI, including through redress.Redress mechanisms thus constitute only one of the essential elements of accountability and this	Chapter III – Assessing Trustworthy AI (page 24)The Draft Guidelines provide an assessment list to operationalise the implementation and assessment of the requirements of trustworthy AI. CIPL welcomes the constructive and operational approach of the Draft Guidelines in this regard. CIPL recommends the final Guidelines make clear that this list should be construed as a tool box where organisations can “pick and choose” based on the specific situation and context as each AI application will require different implementations, safeguards and controls. For this tool box to be efficient, the phrasing should be made clearer, more concrete, precise and practical so that it can actually be used by developers. Questions, such as, “is the system equitable in use?” or “do users have the facility to interrogate algorithmic decisions?” should be avoided and replaced by more evidence-based questions, such as “is there a procedure in place to address data subject requests on the logic of the algorithmic decision?” or “is there a company policy on mandating algorithmic traceability?” CIPL agrees that the assessment of trustworthy AI should be based on a “circular model” of continuous improvement and that it is not a “tick-the-box” exercise or an “execute-once-only process”. The assessment should be revised on a regular basis, at intervals relevant to the specific context and risk. This approach is aligned with the common understanding of the elements of accountability, as explained in CIPL’s comments on the second chapter of the Draft Guidelines (Chapter II – Realising Trustworthy AI) and in CIPL’s accountability papers.With regard to the questions relating to respect for privacy and respect for human autonomy (pages 25-26), CIPL refers to its earlier comments made on the first chapter of the Draft Guidelines (Chapter I – Respecting Fundamental Rights, Principles and Values – Ethical Purpose).	General Comments CIPL is currently exploring the interplay between AI and data protection through its project on “Delivering Sustainable AI Accountability in Practice”. This project aims to provide a detailed understanding of the opportunities presented by AI, its challenges to data protection laws and practical ways to address these issues through best practices and organisational accountability. CIPL published its first report “Artificial Intelligence and Data Protection in Tension” on 10 October 2018. The paper is available here, https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_ai_first_report_-_artificial_intelligence_and_data_protection_in_te....pdf.As further explained in CIPL’s Report, AI is in tension with many long-established data protection principles. It is critical that any set of guidelines addressing AI and the protection of personal data acknowledges such tensions and provides novel, flexible, risk-based and creative approaches to addressing relevant challenges — even if this means departing from conventional interpretations of privacy principles. It will be followed by a second report which will address some of the tools that promote accountability for organisations’ use of AI within existing legal and ethical frameworks, as well as reasonable interpretations of existing principles and laws that will help organisations and regulators to achieve efficient, effective privacy protection in the AI context. CIPL is a global data privacy and cybersecurity think tank in the law firm of Hunton Andrews Kurth LLP and is financially supported by the law firm and 74 member companies that are leaders in key sectors of the global economy. CIPL’s mission is to engage in thought leadership and develop best practices that ensure both effective privacy protections and the responsible use of personal information in the modern information age. CIPL’s work facilitates constructive engagement between business leaders, privacy and security professionals, regulators and policymakers around the world. For more information, please see CIPL’s website at http://www.informationpolicycentre.com/. Nothing in this submission should be construed as representing the views of any individual CIPL member company or of the law firm of Hunton Andrews Kurth.

https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_ai_first_report_-_artificial_intelligence_and_data_protection_in_te....pdf. It addresses in greater detail many of the issues raised in these Draft Guidelines.As a general comment, CIPL supports the pragmatic approach taken in the introduction of the Draft Guidelines. It acknowledges the context-specificity of AI and that "different situations raise different challenges" (page 3 of the Draft Guidelines). This confirms that the implementation of the Guidelines should be tailored to specific use cases. The example provided is indeed very relevant: using AI systems to recommend medical treatment is clearly beneficial to patients and health research more generally. Such AI systems, however, can only provide accurate and adequate insights if they have sufficient health data for a large number and wide variety of patients with similar diseases and who are subject to several medical treatments over a certain period of time. Such a requirement is clearly in tension with an individual's right of erasure under applicable data protection law if the individual's health data is part of these datasets. As a result, while respect for privacy rights must be a key consideration in the development of AI, such rights are not absolute and must be balanced against other human rights and interests (such as those relating to the right to life and health in the previous example) and the benefits of the AI to society as a whole. Any assessment of a particular AI application should therefore consider not only the risks to individuals, but also the benefits to individuals and to society. Additionally, CIPL welcomes the recognition that the Guidelines should aim to foster global reflection and discussion on an ethical framework for AI. There is indeed a need for an international approach to defining the framework as AI develops. Not only are AI applications being used internationally, but the digital supply chain of an AI system often extends beyond borders. Global consistency between AI frameworks is key to providing AI developers and users with legal certainty and assurance that such frameworks can operate efficiently and effectively in global contexts. CIPL stresses that AI cannot be solely developed at a national or regional level and that it often relies on cloud-based technologies. AI is a global technology by nature. Thus, the Guidelines should also take into account more global perspectives on the ethical and responsible development and use of AI and include as one of its objectives the development of international frameworks.Finally, in line with the statement that the Guidelines should be seen as "a starting point for the debate on Trustworthy AI", CIPL recommends the HLEG clarify that the Guidelines should be used as a reference document in Member States' national discussions on AI. Furthermore, CIPL would welcome additional information on the process of updating the Guidelines and how stakeholders' input will be further collected.

appropriate, from public authorities using AI-based automatic scoring systems. 4. Ethical Principles in the Context of AI and Correlating Values (page 8) The Draft Guidelines "lists five principles and correlated values that must be observed to ensure that AI is developed in a human centric manner". While CIPL agrees with the five principles to be observed, it recommends revising the wording that the principles "must be observed" to reflect a more flexible and pragmatic approach that accounts for the variety and fast evolving character of AI technologies. For example, "developers and users should strive to observe the principles". In addition, CIPL believes that a more nuanced approach should be taken to the last three principles (human agency preservation, fairness and transparent operation) for the reasons given further below.CIPL welcomes the Draft Guidelines' recognition of the need to resort to internal and external experts to advise on the design, development and deployment of AI given its "potential of unknown and unintended consequences". In this context, CIPL supports the initiatives of organisations to set up Boards entrusted with advising on AI and Ethics matters.CIPL cautions, however, against the risk of having too many views and possibly conflicting recommendations between experts with different backgrounds on similar questions. As an illustration, CIPL refers to the area of scientific health research where there are often many stakeholders involved in the assessment and review of proposed research projects: data protection authorities (DPAs), health regulatory authorities, patient committees and/or ethical review boards. On the question as to whether a patient should provide consent for the processing of their personal health data for a specific research project, such experts may take different positions. The DPA may consider that consent is not required in such a case on the basis of Article 9(i) or 9(j) of the GDPR whereas a patient committee may deem that consent is required as a key patient right. Diverging views and interpretations on this subject is contributing to a state of legal uncertainty for health research that will ultimately hamper critical health research developments. More information about this topic is available at https://www.huntonprivacyblog.com/2019/01/09/cipl-co-hosts-workshop-on-gdpr-and-scientific-health-research/. When transposed to the AI context, we need to avoid a situation in which too many experts — and, in particular, external experts — provide recommendations or express positions that may contradict one another (or even binding applicable laws, such as the GDPR), without a clear understanding as to which expert(s) (or laws) should have the last say or which position should prevail. Such a state of uncertainty would make it challenging for AI developers and users to navigate the legal and ethical framework applicable to AI. The risk of such inconsistency will even be higher in the context of international projects. As a result, whenever AI relies on the processing of personal data, data protection experts (e.g. legal advisors, Data Protection Officers, data protection committees, DPAs, etc.), informed by applicable data protection laws, should have the ultimate say, taking into account the non-binding positions or recommendations of other experts providing

should be no different in the AI context. One key feature of accountability is that it places the burden of protecting individuals primarily on organisations. When organisations carry out this responsibility effectively, they create trust among the public and regulators that they are processing personal data responsibly, even in the absence of direct individual involvement.Accountability involves (1) setting privacy protection goals based on criteria established in law, self-regulation and best practices; (2) vesting the organisation with both the ability and the responsibility to determine appropriate, effective measures to reach those goals; and (3) having the ability to demonstrate capacity to achieve specified privacy objectives. On this basis, accountability-based data privacy and governance programs typically encompass and address the following elements of accountability: (1) Leadership and Oversight; (2) Risk Assessment; (3) Policies and Procedures (including Fairness and Ethics); (4) Transparency; (5) Training and Awareness; (6) Monitoring and Verification; (7) Response and Enforcement. All the requirements of trustworthy AI provided by the HLEG could be mapped to these seven building blocks of accountability.Accountability is also well-suited to address the ethical risks that have to be identified, assessed and mitigated in the context of AI. Indeed, risk assessment is a core element of accountability. It enables organisations to understand the potential risks and harms to individuals that may be associated with their processing operations, and this, by definition, may include ethical considerations. It also requires organisations to implement appropriate mitigations for such risks and harms, taking into account the desired benefits of the processing and rights and interests of individuals. The accountability approach is consistent with other areas of compliance, including anti-bribery, anti-money laundering, export control and competition. It has been used by organisations, regulators and courts to determine if an organisation has maintained an effective and comprehensive compliance program in a given regulatory area. Organisations are already used to working on this basis and keeping the same structure with similar components for AI technologies, with the necessary adaptations, would enable them to leverage existing and efficient schemes. They could include their programs for accountable AI into their existing data protection accountability programs, which would accelerate trustworthy AI. CIPL calls, therefore, for a common understanding and interpretation of the concept of accountability in the world of AI for all stakeholders — organisations implementing accountability, regulators that are enforcing it and individuals who are seeking effective protection. This issue is discussed in detail in the above-mentioned white papers on organisational accountability. Finally, the understanding of accountability set forth above has become a cornerstone of effective data protection and a dominant trend in global data privacy law, policy and organisational practices. Indeed, the term encapsulates what most regulators now expect of responsible organisations that handle personal data and what many privacy frameworks and data protection laws have incorporated as a matter of basic obligation

guidance on different areas.(a) The Principle of Autonomy: "Preserve Human Agency" (pages 9-10) The Draft Guidelines state that "[i]f one is a consumer or user of an AI system this entails a right to decide to be subject to direct or indirect AI decision making, a right to knowledge of direct or indirect interaction with AI systems, a right to opt-out and a right of withdrawal".While CIPL supports the position that individuals should have the right to know if they interact with an artificial intelligence system, and that the promotion of individual rights and empowerment is an important factor for the development of AI, extensive and unlimited rights to information, to opt-out and to withdrawal may not be appropriate in many instances. Sometimes, the exercise of certain individual rights runs counter to the benefits of certain AI applications. For example, the exercise of the right to erasure may be inappropriate for AI applications where the risk of retaining the data to individuals is low but the deletion of data would prejudice the whole dataset. Such is the case in clinical trials or for scientific health research which carry obvious and huge benefits to society as a whole — enabling one data subject to delete his or her personal data might ultimately cause harm to others as the medication, its prescription or its dosage may not be as accurate as it could have been with more data. A similar argument applies to the ability to "opt-out" of data processing where the processing of personal data is conducted in an accountable fashion to prevent harm to the individual and where opting out would, therefore, serve no data protection objective and would diminish the value of available data sets and undercut legitimate research and product development. Having an overly rigid interpretation of these rights may also prevent public authorities from performing their duties for the common good (e.g. tax collection, social welfare and education). As a consequence, the magnitude of these rights must vary according to the specific data use context.Footnote 13 of the Draft Guidelines, relating to the use of AI in the working environment, states that the principle of autonomy and the rights detailed above include "a right to individually and collectively decide on how AI systems operate in a working environment. This may also include provisions designed to ensure that anyone using AI as part of his/her employment enjoys protection for maintaining their own decision making capabilities and is not constrained by the use of an AI system". It is important to note, however, that there may be instances where full human agency is not workable in practice. For example, AI technology may be very efficient in the context of network security where it can assist in uncovering or even predicting security attacks or incidents affecting the IT resources of a company (whether such incidents come from internal or external sources). Enabling employees to decide (collectively or individually) to exercise their right to opt-out and to withdraw personal data may defeat the whole purpose of an AI system which needs to monitor activity — including employee activity — on the company's IT network. Enabling these rights would, in fact, reduce the relevance of the AI technology and, ultimately, weaken the security of the company network.As a general comment, or best practice (GDPR, OECD Privacy Guidelines, Council of Europe Convention 108, APEC Privacy Framework). Data privacy regulators in numerous jurisdictions have issued regulatory guidance or enforcement orders encouraging or requiring accountability including, Canada, Mexico, Hong Kong, Singapore, Australia, Colombia and the United States. Therefore, using accountability as the architecture for trustworthy AI in the EU would also enable bridge building with other legal regimes outside of Europe.3. Design for all (page 15)While CIPL welcomes the inclusion of this principle to enable all citizens to take advantage of the benefits of AI, it is not realistic to request that every AI system be designed in a way that is designed for all. The Guidelines should take a more nuanced approach and CIPL recommends that the concept to "design for all" be framed as an overall objective that takes into account the particular AI use case, relevance and feasibility of such design.5. Non-Discrimination (page 16)The Draft Guidelines state that "[d]iscrimination in an AI context can occur unintentionally due to, for example, problems with data such as bias, incompleteness and bad governance models...[t]herefore upstream identification of possible bias, which later can be rectified, is important to build in to the development of AI". CIPL agrees with the Draft Guidelines on this point and believes that awareness training and investment in research around identifying and mitigating bias and discrimination is critical to assisting with the upstream identification of bias and discrimination. Furthermore, CIPL suggests that specific guidance and principles addressing discrimination be elaborated. In addition, CIPL recommends the Guidelines recognise that access to and the use of sensitive data can be one method of reducing bias in AI applications. AI technologists have confirmed that in order to avoid bias and discriminatory impacts of AI, algorithms must be tested by reference to sensitive categories of data, such as gender, race and health. Denying access to or preventing the retention of such sensitive data makes it more difficult to detect and remedy bias and may even have the opposite effect of producing biased outcomes without an ability to explain why the AI application is arriving at such discriminatory conclusions. Of course, where sensitive data is processed, appropriate protections, such as masking, security measures, including pseudonymisation, and other accountability safeguards will be of increased importance.7. Respect for Privacy (page 17)Regarding the requirement of respect for privacy, the Draft Guidelines note that "[p]rivacy and data protection must be guaranteed at all stages of the life cycle of the AI system" and that "[o]rganisations must…ensure full compliance with the GDPR as well as other applicable regulation dealing with privacy and data protection".CIPL agrees with this requirement in principle but wishes to highlight a point that it elaborates on in its paper "Artificial Intelligence and Data Protection in Tension" (See https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_ai_first_report_-_artificial_intelligence_and_data_protection_in_te....pdf) — AI is in tension with many long-established data protection principles.

CIPL believes that such "human agency" should be proportional to the actual risk of AI to the rights and freedoms of individuals — a personalised search result on a shopping platform powered by an AI technology is unlikely to harm an individual's right to self-determination — and recommends that this point be specifically mentioned in the Guidelines. This would echo the GDPR's risk-based approach that enables organisations to tailor and calibrate their mitigations and controls to the actual risks presented to the rights and freedoms individuals (see Articles 32 to 36 of the GDPR). Moreover, the appropriate implementation of organisational accountability (including redress procedures) will ensure the effective protection of individuals in contexts where they do not or cannot opt-out of certain data processing.(b) The Principle of Justice: "Be Fair" (page 10)With respect to fairness, the Draft Guidelines note that "[d]evelopers and implementers need to ensure that individuals and minority groups maintain freedom from bias, stigmatisation and discrimination". As a preliminary comment, the Guidelines should acknowledge that in some specific cases, biased results may be intended because of the particular objective of the AI system (e.g. AI analysis for the purpose of positive discrimination).Generally, discrimination and unfairness in the AI context may occur unintentionally because AI systems are fed with datasets that may themselves carry some kind of bias. AI technologists have confirmed that in order to avoid bias and discriminatory impacts of AI, algorithms must be tested by reference to sensitive categories of data, such as gender, race and health. Denying access to or preventing the retention of such sensitive data makes it more difficult to detect and remedy bias and may even have the opposite effect of producing biased outcomes without an ability to explain why the AI application is arriving at such discriminatory conclusions. Therefore, CIPL recommends that access to and the use of such sensitive data be facilitated as one method of reducing bias in AI applications. Of course, where sensitive data is processed, appropriate protections, such as masking, security measures, including pseudonymisation, and other accountability safeguards will be of increased importance.(c) The Principle of Explicability: "Operate Transparently" (page 10)With respect to the transparent operation of AI technologies, the Draft Guidelines state that "[t]ranspareny is key to building and maintaining citizen's trust in the developers of AI systems and AI systems themselves" and AI systems should be "comprehensible and intelligible by human beings at varying levels of comprehension and expertise". CIPL agrees that transparency is an essential factor in generating and maintaining trust in AI applications and that taking into account the different levels of transparency and information required for each relevant audience is a key consideration in improving the transparency and intelligibility of AI systems. However, CIPL stresses that defining the right level of transparency of information on the functioning of AI systems may be challenging. CIPL further cautions against a broad interpretation of this principle that may have the unintended consequences of (1) stifling innovation because of the potential access to strategic

It is critical that any set of guidelines addressing data protection in the context of AI acknowledges such tensions and provides novel, flexible, risk-based and creative approaches to addressing relevant challenges — even if this means departing from conventional interpretations of privacy principles. Furthermore, while respect for privacy rights must be a key consideration in the development of AI, it must be recalled that such rights are not absolute and must be balanced against other human rights and also with the benefits that AI could bring to society as a whole. Finally CIPL underlines the huge potential for AI to further enhance privacy and calls for further work in this area. AI can help, for instance, in building robust anonymisation techniques to make sure that personal data cannot be re-identified.8. Robustness (page 17)As part of the robustness principle, the Guidelines describe accuracy as "AI's confidence and ability to correctly classify information into the correct categories, or its ability to make correct predictions, recommendations, or decisions based on data or models". CIPL underlines that, in practice, the level of acceptable accuracy for a specific AI system will vary depending on the risk and would need to be defined on a case-by-case basis.CIPL welcomes the Draft Guidelines proposal for a fall back plan in case of problems with the AI system, such as switching from statistical to rule-based procedures or asking for a human operator before continuing the action. CIPL recommends further exploring this concept, keeping in mind that its implementation and operation will vary depending on the specific AI use case and technical feasibility.10. Transparency (page 18)The Draft Guidelines state that "[b]eing explicit and open about choices and decisions concerning data sources, development processes, and stakeholders should be required from all models that use human data or affect human beings or can have other morally significant impact".CIPL supports this statement in principle, but believes that such a requirement, without qualification, is too far reaching and recommends a more nuanced approach with respect to the transparency requirement in the final Guidelines.Transparency is one of the key building blocks of accountability that is instrumental in fostering trust in the development and use of AI systems. However, the objective of reduction of information asymmetry between organisations and individuals should be context specific and balanced with other factors, such as the protection of commercially-sensitive information, the risk posed by the AI system to individual and the availability of transparency in redress mechanisms. The transparency standard for AI should also take into account existing laws such as the GDPR when personal data is processed. CIPL recommends the final Guidelines take such considerations into account.For more information on CIPL's views concerning transparency, we refer you to our earlier comments in relation to Chapter I – Respecting Fundamental Rights, Principles and Values – Ethical Purpose.2. Technical and Non-Technical Methods to Achieve Trustworthy AI (page 18)The Draft Guidelines state "given that AI systems are continuously evolving and acting in a dynamic environment, achieving Trustworthy

competitive information or trade secrets and (2) overloading individuals with information that may not be understandable or useful to them. Moreover, some low risk AI-based processing may require less "transparency" or information to individuals to avoid overwhelming them with information. Thus, the level of involved residual risk despite appropriate mitigations and controls should be a consideration in devising the appropriate level and nature of "transparency".The Draft Guidelines state that explicability provides for the possibility of individuals and groups to request "evidence of the baseline parameters and instructions given as inputs for AI decision making". CIPL wishes to highlight that Article 13(2)(f) of the GDPR already requires that "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" be provided to individuals in the context of automated decision-making and profiling. CIPL recommends that a pragmatic approach to "algorithmic transparency" should be based on a broad understanding of "logic involved" and should focus on a useful and actionable level of transparency (including information on whether decisions are automated, what factors they are based on and, where relevant, information regarding the specific algorithmic logic) coupled with appropriate safeguards. These safeguards can include the right to contest the decision or ask for human review of the decision-making if it results in a material negative impact. Of course, not every decision should be subject to scrutiny or human review (e.g. being presented with an ad for a red car instead of a blue one), but only those that create a legal effect or harm for individuals (e.g. in the context of insurance, employment and credit). To read more about CIPL's recommendations with respect to automated decision-making under the GDPR, please see Comments by the Centre for Information Policy Leadership on the Article 29 Data Protection Working Party's "Guidelines on Automated Individual Decision-Making and Profiling", available at: https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_comments_to_wp29_guidelines_on_automated_individual_decision-making_and_profiling.pdf.It is also important to consider that algorithms cannot be understood in a static manner. Datasets, data models and algorithms all constantly change based on accumulated knowledge and insights. This makes it difficult to deliver real-time and detailed transparency on their workings. Moreover, CIPL cautions that transparency, in the context of AI, may need to be understood in new ways with respect to decisions made by complex AI algorithms. This is partly attributed to the "black box" problem which, in the current state of the art, can make it practically impossible to explain why certain complex algorithms arrive at a specific result. Furthermore, for certain AI applications, the technology involved will be too complex and the user of the technology may not have the means to provide complete information to the individual.CIPL recommends that the Guidelines acknowledge these difficulties and include additional options that can deliver meaningful information and empowerment of the individual. This could include human review of AI decisions, redress mechanisms, AI is a continuous process". In effect, this means that requirements and methods for accountable AI must be regularly evaluated and updated through proactive, ongoing monitoring, periodic risk assessments and other accountability measures. The accountability-based approach to data protection and data governance is uniquely suited to this iterative nature of compliance in the AI context and should be emphasised more explicitly in the final Guidelines.CIPL has worked extensively on the topic of accountability and has published several papers on the central role of organisational accountability in data protection. To read more about CIPL's work in this area please see the following papers: "Introduction: The Central Role of Organisational Accountability in Data Protection", available at https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/introduction_to_two_new_cipl_papers_on_the_central_role_of_organisational_accountability_in_data_protection.pdf; "The Case for Accountability: How it Enables Effective Data Protection and Trust in the Digital Society", available at https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_accountability_paper_1_-_the_case_for_accountability_-_how_it_enables_effective_data_protection_and_trust_in_the_digital_society.pdf and "Incentivising Accountability: How Data Protection Authorities and Law Makers Can Encourage Accountability", available at https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_accountability_paper_2_-_incentivising_accountability_-_how_data_protection_authorities_and_law_makers_can_encourage_accountability.pdf.With respect to non-technical methods to ensure trustworthy AI, CIPL recommends adding "global governance" to the final Guidelines. AI is often developed, deployed and used across borders as part of a wider ecosystem. As a result, on-going dialogue and the promotion of common norms across geographies are key to the development of accountable AI. The Draft Guidelines include codes of conduct as one non-technical method to ensure trustworthy AI. CIPL fully supports the reliance on co-regulation tools such as certifications and codes of conduct in this respect. Article 24(3) of the GDPR recognises the importance of approved codes of conduct and certification mechanisms for the purpose of demonstrating accountability. Such schemes can provide a framework for organisations while giving assurances to individuals or other organisations that an accredited third party has reviewed and approved the processing at issue in a particular AI context. They can play an important role in ensuring the accountability of AI systems and should be further promoted. There is an urgent need, particularly in the EU, to build the regulatory framework that enables the development of both of these accountability tools (certifications and codes of conduct) and to ensure these schemes are designed to be applicable to AI and machine learning contexts. As far as standardisation is concerned, in line with its previous comments, CIPL recommends against setting up standards across applications and sectors that may be too rigid to capture the diversity and fast evolving nature of AI.

feedback tools and specific solutions for more sophisticated audiences, i.e. regulators. Additionally, employing security measures, such as pseudonymisation and anonymisation, where appropriate and feasible, can help compensate for any obstacles to providing full transparency to individuals.5. Critical Concerns Raised by AI (page 11) The AI HLEG puts forward identification without consent as one of the critical concerns raised by AI in the Draft Guidelines (Section 5.1) The Guidelines note that "[d]ifferentiating between the identification of an individual vs. the tracing and tracking of an individual, and between targeted surveillance and mass surveillance, will be crucial for the achievement of Trustworthy AI" and that addressing this concern "involves an ethical obligation to develop entirely new and practical means by which citizens can give verified consent to being automatically identified by AI or equivalent technologies".In effect, this means that without obtaining the consent of an individual, identification cannot occur and all services derived from such identification cannot be offered. While CIPL appreciates that the Draft Guidelines recognise that the identification of individuals is, in some instances, aligned with ethical principles (e.g. for fraud detection and prevention, detecting money laundering or terrorist financing), CIPL wishes to underline some of the key consequences of requiring that consent for identification be given in instances where the particular use of AI is not clearly warranted by existing law or by the protection of core values:Mandating consent as a general rule for identification except in the limited situations outlined above would run counter to the GDPR — the Draft Guidelines correctly note that Article 6 of the GDPR provides that processing of data shall only be lawful if it has a valid legal basis. There are six possible legal bases under Article 6 (consent, contractual necessity, compliance with a legal obligation, protection of the vital interest of the data subject, public interest or legitimate interest of the controller or a third party). These legal bases are all placed on equal footing under the GDPR. No one basis is privileged over another. Under the current reading of the Draft Guidelines, Article 6(1)(a) would become the norm for the identification of individuals in the context of AI, with the other legal bases becoming de facto moot. This is certainly not the intent of the GDPR. In addition, the GDPR does not make a distinction between different technologies and does not empower data subjects to object to the use of a specific technology itself. In addition, overreliance on consent undermines its quality and creates consent fatigue for individuals who are increasingly using digital services and technology in their daily personal and professional lives. In other words, excessive reliance on consent does not put individuals in true control as when faced with a large number of consent requests, consumers may resort to clicking through without understanding what exactly they are consenting to. The HLEG has acknowledged this by stating "[a]s current mechanisms for giving informed consent in the internet show, consumers give consent without consideration".Moreover, consent may not be the most appropriate legal basis in the AI world as it may not even be possible to achieve the requirements for

valid consent provided for by the GDPR. Under Article 4(11), consent must be freely given, specific, informed, unambiguous and given by a clear affirmative action. As a result, consent should be used as a legal ground for processing in the AI context only where it can be obtained in line with all the GDPR requirements and where it is meaningful. This would require that the following conditions are fulfilled: (a) it is possible to provide clear and understandable information; (b) individuals have a genuine choice to decide whether to use a service or not; (c) consent is not used in situations where there is a clear imbalance in the relationship between the organisation and the individuals and (d) consent can be withdrawn without any detriment to individuals. Moreover, even where consent would be feasible, where other legal bases apply, organisations should be able to rely on them. Indeed, other bases may be more protective of individuals, such as the legitimate interest basis, as discussed in more detail in CIPL's paper on "Recommendations for Implementing Transparency, Consent and Legitimate Interest under the GDPR", available at https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_recommendations_on_transparency_consent_and_legitimate_interest_under_the_gdpr_-19_may_2017-c.pdf. The legitimate interest legal ground is particularly useful because it can provide a valid basis in a wide range of situations where the processing of personal data through AI systems has little to no impact on individuals or does not create any risks (or if there are risks, this ground requires appropriate mitigations to reduce or remove them). Moreover, identification, even if automatic, may not necessarily carry high risk for the individual. Examples of where AI systems provide benefits and where the processing of data is based on legitimate interest are wide ranging. They include spam and fraud prevention, improvements in healthcare provision and disease prevention, environmental protections, scientific advancement, timely payment processing and invoicing, cybersecurity or tax collection. Furthermore, relying on consent in the context of AI may be problematic as the GDPR also provides individuals with a right to withdraw consent. The exercise of such a right can be detrimental to the relevance and reliability of the dataset which feeds the AI system. In fact, the mere possibility to withdraw consent itself does not provide sufficient legal certainty and trust in the datasets used and may hinder some AI systems' development.Thus, while CIPL supports the Draft Guidelines' call for developing new means by which verified and meaningful consent can be provided, in many AI contexts — including the use of AI for identification purposes — it may be more appropriate for organisations to rely on other legal grounds for processing, such as legitimate interest or contractual necessity. Finally, individuals will not want to be burdened with approving every single use of data or every identification of themselves or every single processing operation necessary for the provision of products and services that they want to use. In fact, individuals will expect organisations to use data and develop products, services and technology in a responsible and accountable manner and

to use consent as a means to legitimise data use in situations where there is a clear and easy choice for individuals. Indeed, in cases where the use of consent may not be available, realistic or practicable, there are other tools that can protect the individual. Examples of such tools and concepts include transparency, risk assessments, alternative legal bases, including the legitimate interest balancing test, organisational accountability, data protection by design, security measures, exercise of individuals' rights, redress in cases of infringement, etc. Unlike consent, these tools provide the particular speed, scale and flexibility necessary to serve the demands of AI technologies. These other protections are inherent in the concept of organisational accountability, as described above and in our recent white papers on the topic. See "Introduction: The Central Role of Organisational Accountability in Data Protection", available at https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/introduction_to_two_new_cipl_papers_on_the_central_role_of_organisational_accountability_in_data_protection.pdf; "The Case for Accountability: How it Enables Effective Data Protection and Trust in the Digital Society", available at https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_accountability_paper_1_-_the_case_for_accountability_-_how_it_enables_effective_data_protection_and_trust_in_the_digital_society.pdf and "Incentivising Accountability: How Data Protection Authorities and Law Makers Can Encourage Accountability", available at https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_accountability_paper_2_-_incentivising_accountability_-_how_data_protection_authorities_and_law_makers_can_encourage_accountability.pdf.

Angeles	Manjarrés	Universidad Nacional de Educación a Distancia, Spain			Throughout this chapter, there is not enough emphasis on warning of possible psychological damage, on how intelligent technology could negatively influence mental health and personal fulfillment, that is to say living a life with meaning and purpose. No mention is made of major problems associated with the use of this technology, such as the alteration of the concept of identity and the nature of human interactions, the blurring of the distinction between the real and the virtual, escapism to virtual worlds, the substitution and deterioration of human bonds, cognitive overload, the loss of meaning and purpose due to being replaced by intelligent machines ... Nor the potential loss of human values: wisdom, creativity, empathy, affection, social skills ...The "Target audience" should clearly include teachers, since to have any real impact, ethics guidelines need to be an integral part of the education of future AI practitioners, in particular, in university-level education. In fact, in our opinion, the importance of ethics guidelines in university-level AI education merits specific treatment somewhere in the document.The document presents a rather Eurocentric view, in contradiction to the European commitment to the UN 2030 Agenda. Although the Sustainable Development Goals (SDGs) are mentioned in the document, they are not given sufficient importance. The primary objective of putting technology at the service of an equitable world is not highlighted. In particular, Europe is committed to addressing the eradication of poverty as well as the refugee crisis through humanitarian assistance and civil protection actions, as well as through international cooperation for development. (see, for example, the following documents:- European Commission Fact Sheet "Next steps for a sustainable European future - European action for sustainability: Questions & Answers",http://europa.eu/rapid/press-release_MEMO-16-3886_en.htm- European Commission, "10 Commission priorities for 2015-19", num. 9 "A stronger global actor"https://ec.europa.eu/commission/priorities/stronger-global-actor_en)These commitments should be reflected in the document. Currently, there are just a few references to "vulnerable groups" or "asymetries of power or information". The potential of Artificial Intelligence applications for contributing to addressing global challenges such as climate change, lack of high quality services to excluded populations, poverty, exploitation, violations of human rights and increased violence, the world-wide refugee crisis, hunger, etc. should be highlighted. It has been demonstrated that AI technologies can make a significant contribution to achieving the UN SDGs, through the development of fields such as "big data for development" (applications in agriculture, medical tele-diagnosis,...); geographic information systems (applications in public service planning, disaster prevention, emergency planning, disease monitoring, improving refugee resettlement); control systems (applications in naturalizing intelligent cities through energy and traffic control, management of urban agriculture), etc. Testimony to this fact, for example, are the actions of United Nations Global Pulse (see, for example, "Big Data for Development: Challenges and Opportunities", 2012). It is	The interest of having AI systems that include, by design, modules to facilitate the collection of data for the calculation of impact indicators concerning the systems themselves, as well as providing support for this calculation, should be made clear. These indicators should cover a broad spectrum (all kinds of well-being indicators).In the case of applications developed for LMICs, it is advisable to include experts in "technologies for development" and "development studies", National and International Government Agencies and NGOs with expertise in technology sectors, as part of the multidisciplinary team participating in AI system development.In the case of LMICs, special attention should be paid to implementation and deployment difficulties, particularly adaptation to the available resources (hardware, software, connectivity ...), the impact on the receiving communities, the suitability and sustainability of the applications in all dimensions ...Emphasis should be placed on collecting data for a broader concept of the impact measure, associated with compliance with the SDGs, including social costs, impact on the workplace, and taking into account the values of the culture in which they are delivered. In the case of applications focused on development objectives in LMICs, attention should be drawn to indicators related to specific priorities and those typically used in cooperation for development actions (schooling, average life expectancy, access to basic services ...).		This contribution has two authors, Angeles Manjarrés, whose details are given above and Simon PickinEmail address: simon.pickin@fdi.ucm.esOrganization: Universidad Complutense de Madrid, SpainThe authors are participating in "The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems", in particular, we have collaborated in writing the report "Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems ", version 2 and version 3 (which is scheduled to be released in February). In particular, we have collaborated in the chapter entitled "A / IS for Sustainable Development and a More Equal World" (EADv3), formerly entitled "Economics and Humanitarian Issues" (EADv2). We provide a summary of the content that we think is missing from the Trustworth AI document with respect to the work done in the IEEE project, in particular, concerning the aforementioned chapter.

of interest to point out that these ethical considerations pertain to a wider view of ethics, focusing on potentialities and not only on risk mitigation, macro-ethics rather than micro-ethics.It is also important to draw attention to the idiosyncrasy of the development of applications in the LMICs (Lower and Middle Income Countries) context and the particular impact that these technologies can have in these countries. In this regard, it would be useful to point out the relevance, among other things, of:- Particular aspects to which attention must be paid in the case of assistance in humanitarian crises, see, for example, the following documents:* 10 big data science challenges facing humanitarian organizationshttps://www.unhcr.org/innovation/10-big-data-science-challenges-facing-humanitarian-organizations/.* The Signal Codehttps://signalcode.org/- Culture-aware principles, see, for example, the following document:* "Culturally-Aware HCI Systems":https://link.springer.com/chapter/10.1007/978-3-319-67024-9_2- It would be of interest to say something about "Open AI", where this refers not only to FOSS (Free / Open-Source Software) but also to applying FOSS principles to algorithms, scientific insights or other AI artifacts. See, for example the following documents/initiatives:* Strategic Implications of Openness in AI Development. Nick Bostrom:https://onlinelibrary.wiley.com/doi/epdf/10.1111/1758-5899.12403* Open Source AI For Everyone. The Linux Foundation:https://www.linuxfoundation.org/blog/2018/05/open-source-ai-for-everyone-three-projects-to-know/* OpenAI research company:https://openai.com/about/* Google, Advancing AI for Everyone:https://ai.google/Also of interest is the relation between OpenAI and the principle of "explicability"There is no warning about the possibilities of control, manipulation, attack on autonomy, etc. raised by the field of affective computing, given the susceptibility of humans to emotional influence. The specificities of this problem require separate treatment and particular methodological considerations. Neither does the document highlight the potentially important effects on mental and physical health that virtual immersive interactive applications could have (with intensive application of Artificial Intelligence techniques).

Angeles	Manjarrés	Universidad Nacional de Educación a Distancia, Spain		Throughout this chapter, there is not enough emphasis on warning of possible psychological damage, on how intelligent technology could negatively influence mental health and personal fulfillment, that is to say living a life with meaning and purpose. No mention is made of major problems associated with the use of this technology, such as the alteration of the concept of identity and the nature of human interactions, the blurring of the distinction between the real and the virtual, escapism to virtual worlds, the substitution and deterioration of human bonds, cognitive overload, the loss of meaning and purpose due to being replaced by intelligent machines ... Nor the potential loss of human values: wisdom, creativity, empathy, affection, social skills ...The "Target audience" should clearly include teachers, since to have any real impact, ethics guidelines need to be an integral part of the education of future AI practitioners, in particular, in university-level education. In fact, in our opinion, the importance of ethics guidelines in university-level AI education merits specific treatment somewhere in the document.The document presents a rather Eurocentric view, in contradiction to the European commitment to the UN 2030 Agenda. Although the Sustainable Development Goals (SDGs) are mentioned in the document, they are not given sufficient importance. The primary objective of putting technology at the service of an equitable world is not highlighted. In particular, Europe is committed to addressing the eradication of poverty as well as the refugee crisis through humanitarian assistance and civil protection actions, as well as through international cooperation for development. (see, for example, the following documents:- European Commission Fact Sheet "Next steps for a sustainable European future - European action for sustainability: Questions & Answers",http://europa.eu/rapid/press-release_MEMO-16-3886_en.htm- European Commission, "10 Commission priorities for 2015-19", num. 9 "A stronger global actor"https://ec.europa.eu/commission/priorities/stronger-global-actor_en)These commitments should be reflected in the document. Currently, there are just a few references to "vulnerable groups" or "asymetries of power or information". The potential of Artificial Intelligence applications for contributing to addressing global challenges such as climate change, lack of high quality services to excluded populations, poverty, exploitation, violations of human rights and increased violence, the world-wide refugee crisis, hunger, etc. should be highlighted. It has been demonstrated that AI technologies can make a significant contribution to achieving the UN SDGs, through the development of fields such as "big data for development" (applications in agriculture, medical tele-diagnosis,...); geographic information systems (applications in public service planning, disaster prevention, emergency planning, disease monitoring, improving refugee resettlement); control systems (applications in naturalizing intelligent cities through energy and traffic control, management of urban agriculture), etc. Testimony to this fact, for example, are the actions of United Nations Global Pulse (see, for example, "Big Data for Development: Challenges and Opportunities", 2012). It is	The interest of having AI systems that include, by design, modules to facilitate the collection of data for the calculation of impact indicators concerning the systems themselves, as well as providing support for this calculation, should be made clear. These indicators should cover a broad spectrum (all kinds of well-being indicators).In the case of applications developed for LMICs, it is advisable to include experts in "technologies for development" and "development studies", National and International Government Agencies and NGOs with expertise in technology sectors, as part of the multidisciplinary team participating in AI system development.In the case of LMICs, special attention should be paid to implementation and deployment difficulties, particularly adaptation to the available resources (hardware, software, connectivity ...), the impact on the receiving communities, the suitability and sustainability of the applications in all dimensions ...Emphasis should be placed on collecting data for a broader concept of the impact measure, associated with compliance with the SDGs, including social costs, impact on the workplace, and taking into account the values of the culture in which they are delivered. In the case of applications focused on development objectives in LMICs, attention should be drawn to indicators related to specific priorities and those typically used in cooperation for development actions (schooling, average life expectancy, access to basic services ...).		This contribution has two authors, Angeles Manjarrés, whose details are given above and Simon PickinEmail address: simon.pickin@fdi.ucm.esOrganization: Universidad Complutense de Madrid, SpainThe authors are participating in "The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems", in particular, we have collaborated in writing the report "Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems ", version 2 and version 3 (which is scheduled to be released in February). In particular, we have collaborated in the chapter entitled "A / IS for Sustainable Development and a More Equal World" (EADv3), formerly entitled "Economics and Humanitarian Issues" (EADv2). We provide a summary of the content that we think is missing from the Trustworth AI document with respect to the work done in the IEEE project, in particular, concerning the aforementioned chapter.

				of interest to point out that these ethical considerations pertain to a wider view of ethics, focusing on potentialities and not only on risk mitigation, macro-ethics rather than micro-ethics.It is also important to draw attention to the idiosyncrasy of the development of applications in the LMICs (Lower and Middle Income Countries) context and the particular impact that these technologies can have in these countries. In this regard, it would be useful to point out the relevance, among other things, of:- Particular aspects to which attention must be paid in the case of assistance in humanitarian crises, see, for example, the following documents:* 10 big data science challenges facing humanitarian organizationshttps://www.unhcr.org/innovation/10-big-data-science-challenges-facing-humanitarian-organizations/.* The Signal Codehttps://signalcode.org/- Culture-aware principles, see, for example, the following document:* "Culturally-Aware HCI Systems":https://link.springer.com/chapter/10.1007/978-3-319-67024-9_2- It would be of interest to say something about "Open AI", where this refers not only to FOSS (Free / Open-Source Software) but also to applying FOSS principles to algorithms, scientific insights or other AI artifacts. See, for example the following documents/initiatives:* Strategic Implications of Openness in AI Development. Nick Bostrom:https://onlinelibrary.wiley.com/doi/epdf/10.1111/1758-5899.12403* Open Source AI For Everyone. The Linux Foundation:https://www.linuxfoundation.org/blog/2018/05/open-source-ai-for-everyone-three-projects-to-know/* OpenAI research company:https://openai.com/about/* Google, Advancing AI for Everyone:https://ai.google/Also of interest is the relation between OpenAI and the principle of "explicability"There is no warning about the possibilities of control, manipulation, attack on autonomy, etc. raised by the field of affective computing, given the susceptibility of humans to emotional influence. The specificities of this problem require separate treatment and particular methodological considerations. Neither does the document highlight the potentially important effects on mental and physical health that virtual immersive interactive applications could have (with intensive application of Artificial Intelligence techniques).			
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Emanuela	Girardi	Pop Ai - CLAIRE Supporter	When I think about technology, I always think about something positive and useful that can improve our quality of life. I do not feel trust for a technology, which I believe doesn’t have the attribute “trust”. I need to trust who designs, develops and deploys the technology. I do understand the transformative and disruptive impact that AI will have on people's lives, and I believe that the majority of people are not aware of it. When you do not know something, you are often afraid of it. It is then a matter of awareness and knowledge more than trust. Promoting the concept of a trustworthy AI	Fundamental Rights Among the fundamental rights of human being, I would also add the right to receive a digital education to enable people to participate in the digital society. Concern A severe concern that should be added is represented by the advance of genomics thanks to AI technologies. If we consider for instance CRISPR, the gene-editing technology, it can have a huge positive impact on genetic diseases but it also raises ethical, moral and legal questions related to	Realizing a trustworthy AI in line with the proposed guidelines could limit the development of AI applications by startups and small companies. The guidelines could be followed by big organizations and corporations while small organizations might lack skills and resources to be compliant. This might limit the development of innovative AI applications by European startups.	I think that the 4 proposed use cases of AI will be very useful to understand how to apply the assessment of the AI system in a specific context.	The guidelines are very well formulated, and they will be very useful for the European and the global AI landscape.

			implies that AI could be "not trustable" and could raise negative attitudes towards AI. I would have preferred a human-centric AI more than a trustworthy AI.	the “optimization” of the human race. AGI It is very difficult to regulate something that is not there yet. Today we can just make hypothesis on how artificial consciousness will work and we should try to keep, if possible, humans in the loop. But at the current status of AGI research, the topic is too uncertain to be addressed.			
Lotta	Engdahl	Ministry of Transport and Communication	The Comission’s High-Level Expert Group on Artificial Intelligence (AI HLEG) published in December 18th 2018 a draft of the AI Ethics guidelines for Trustworthy AI. The final version is due in March 2019. We welcome the High-level group’s work as guidance on the concrete implementation and operationalisation into AI systems presented in the working document. The document is a starting point on guidelines, that should be regularly updated over time to ensure continuous relevance as the technology and our knowledge thereof evolves. The document should therefore be considered as the first step for the discussion on “Trustworthy AI made in Europe”.Artificial intelligence (AI) are largely based on the efficient use of data. The development of the data economy increases growth, employment, innovation and the human-centred focus of the global data economy. The goal must be to build a competitive and human-driven data economy in the EU which must be based on extensive availability and use, while respecting the rights and privacy of individuals. The practices of human-centred data management (MyData) shall be strengthened so that individuals have the ability to share and enable efficient use of data for better and individual services. Business should be enabled by creating shared principles, practices and rules for sharing data and granting access rights. This should concretely be delivered through the process. The cooperation on AI intelligence is not only important on EU level, but also important on global level. Europe has to seize the opportunity for trust-based business development. This is also our strength to stand out from other global players in AI. In data economy we need to move our focus from business of scale to business of skills. Consumers rely on big companies and businesses. Therefore we should enforce and strengthen the possibilities for small business to enter the market. Individual trust can be strengthened by adapting shared requirements, principles and understanding of the functionalities of businesses.Transparency serves as a basis for the reliability of an artificial intelligences decision-making. Reliability can be achieved by enabling an external entity to review and evaluate the basis for descision-making. Also, transparency of the algorithms used by AI should be ensured. Only then we can provide authorities with full understanding and make it possible for them to influence the criteria on which the decision-making system is based on, as well as to increase trust amongst individuals. Global markets makes sure that automation is one of the central means used in future global data economy as AI systems uses automated data to achieve the given goal or action. Automation also helps to improve the safety and efficiency of transport and the fulfilment	We share the view that AI holds the promise to increase human wellbeing and the common good, but to do this it needs to be human-centric and respectful of fundamental rights. A digital society must be built on solid trust. When data is used as a raw material for services, it is important to increase trust among users towards products and services. Realising trustworthy AI means individuals right to transfer their data on one service to another in a usable format. Therefore human-centred data economy should be strengthened by creating concrete mechanisms for interoperable structures for data transfer.According to our view, the five ethical principles in chapter I on ethical principles in the context of AI and correlating values (do good, do no harm, preserve human agency, be fair and operate transparently) mentioned, is supported. For citizens the most valuable value is ‘do not spy’, which should be added in the list of values. Strong emphasis should be put on transparency. For example, transparency of the algorithms used by automated vehicles should be ensured. This would provide authorities full understanding and make it possible for them to influence the criteria on which the decision-making of the vehicles’ automated driving system is based, as well as to increase trust amongst road users.	We see the requirements of trustworthy AI listed in the chapter II as a good start to realise the impact of AI in the future. These requirements are:1. Accountability 2. Data Governance 3. Design for all 4. Governance of AI Autonomy (Human oversight) 5. Non-Discrimination 6. Respect for (& Enhancement of) Human Autonomy 7. Respect for Privacy 8. Robustness 9. Safety 10. Transparency The listed requirements are supported and a concrete start when developing, deploying and using AI. Nevertheless the list should also include access to data as one of the requirements. Access to data is the key enabler of AI. Market development should be guided towards decentralized data solutions where data is managed by data access and identity management to be used in digital service networks. Data should be opened on fair, reasonable and non-discriminatory terms. True personalization of services can be achieved by building data transparency and data flows via decentralized data solutions (like blockchain) that connect data economy systems ensuring interoperability and usability of data while preserving data protection and privacy. Blockchain offers also tracybility in terms of to the automated decision-making. This should be adressed thorougfully in the chapter II Non technical methods. It is important that the data gathered is available for use by other services and service providers. For example in transport market the requirements on access to data about routes, parking, scedules and pricing, should be available for other service providers. Also vehicle data should be available for neccesity extent. Therefore interoperability between for example the transport services should be opened. Access to such information should be provided through an interface in the data system. The storing and sharing data should be based on so called on-stop shop principle, i.e. storage of data in one system only should be made possible in order to enable access rights to parties wich have the right to access such data. Infrastructuring a digital twin from data gathered by vehicles (for example weather and road conditions) should be available and opened. The opening of interfaces should be done in a technologically neutral manner.	Autonomous driving/movingTo select autonomous driving/moving as a use case of AI in the draft of this ethics guidelines is very much encouraged in the path of deploying the AI system in the future. We strongly recommend to use the term Autonomous Transport instead of Autonomous driving/moving. Autonomous transport is one of the first fields where AI solutions will be implemented in real life. For example shuttle buses and high lane buses are early adapted means. Artificial intelligence is a key to ensuring the ambitious goals to continue to advance sustainable and safe autonomous traffic in different modes of transport. Data generated for and from the use of automated vehicles as well as collected from traffic, especially by vehicle manufacturers, should be made accessible for the use of third parties and authorities in order to ensure the fluidity, compatibility and safety of automated traffic. As already mentioned in chapter II 1. requirements of trustworthy AI transparency of the algorithms used by automated vehicles should be ensured in order to provide authorities with full understanding and make it possible for them to influence the criteria on which the decision-making of the vehicles’ automated driving system is based, as well as to increase trust among road users. According to our view, the ten requirements of Trustworthy AI are all important to support the operationalization of a further AI. We would like to address a specific glance on the means to prevent faults and wrongly made decisions by AI, instead of paying to much focus on the accountability and monetary compensation mechanisms. When it comes to complex AI systems such as autonomous driving, auditability of the AI system becomes critical and it is essential to have pre-explanationed the methods, logic and the decisions it makes. When the functioning of AI systems can be evaluated against a certain set of criteria and simulated ex ante, then its proper functioning can also be evaluated ex post.	For further discussion, there is a need for frameworks to identify the different levels and to bring up the essential issues related to development of AI. It is important to evaluate the development of AI on a general level and sector specific-level paralelly. The development should be considered together to minimize the gap between regulations on these levels.The level of ambition should be raised by building a sustainable, competitive and human-driven data economy in the EU, which must be based on extensive availability and use of data, while respecting the rights and privacy of individuals. All new initiatives should also take into account the horizontal interoperability and free flow of data in the single market. Business should be enabled by creating common principles, practices and rules for data sharing and access. If these lighter means are not enough, development can be promoted through regulation.Promoting AI and automation is one of the central means of improving the competitiveness of the EU as part of the global data economy. Technological innovations, such as the Internet of Things and AI, enable the collection and use of ever-growing data volumes and the building of key societal services (e.g. smart traffic, healthcare, trade and insurance) on top of this data. We need to strengthen the trust of consumers in technologies, devices and applications related to the safe processing of data. In the societal context transparency of algorithms is necessary to enable discussion about the influence of the decisions and effects of machines on people’s lives and thereafter if necessary regulation is in order (e.g. traffic safety of autonomous transport).

		of environmental and climate goals in all modes of transport. Therefore AI also has a huge meaning in a broad societly context.Automation also applies to the field of humanities, because if you want to strengthen citizens' trust, then action must also have acceptable and solid reasoning. Individuals have poor prospects for anticipating future development and therefore a change in the operating environment requires extensive research and updating of skills. Human behavior also changes in a new situation. In addition to researching artificial intelligence, ethical issues should also be systematically in focus of research and studies.				
Anonymous	Anonymous	Anonymous				Representatives of the Center for Human-Compatible AI have put together a feedback document which can be found here https://docs.google.com/document/d/1mWDmcHh1gwcwZ5OqoJUVILgoGCl0NKTepWWEafXJJXM/edit?usp=sharingIt will be much easier to read via the document at that link, but I've also copied the content below.As representatives of the Center for Human-Compatible AI (CHAI) at UC Berkeley, we welcome the opportunity to feedback on the European Commission's AI HLEG Draft AI Ethics Guidelines For Trustworthy AI.CHAI's mission is to develop the conceptual and technical wherewithal to reorient AI research towards provably beneficial systems, and we tend to focus on longer-term AI outcomes. We therefore consider ourselves to be well-placed to give feedback on section I 5.5 in particular, and would be happy to help revise it.Feedback on section 5.5: Potential longer-term concerns:We agree with the assessment that long-term concerns from advanced artificial could be potentially very high impact. While active regulation in this area is premature, we strongly endorse continued technical research in this area and risk assessment on an ongoing basis.We strongly disagree with some of the characterizations of the concern in this report. In particular, we suggest considering the broader notion of transformative AI: an AI system that precipitates a transition comparable to the industrial revolution. An artificial general intelligence -- a system that can perform any human task -- would certainly be a transformative AI. However, generality is not needed to have a large impact on society: an AI that is superhuman at surveillance and social persuasion would be both a strategic asset and potential threat, even if it were unable to perform human tasks such as driving, singing or dancing. In fact, even superhuman ability in a single narrow domain, such as scientific research, could be enough to transform society.The report mentions some examples of longer-term concerns: "Artificial Consciousness, i.e. AI systems that may have a subjective experience, of Artificial Moral Agents or of Unsupervised Recursively Self-Improving Artificial General Intelligence (AGI)". We would suggest clarifying this statement in two ways:1. Acknowledging that artificial consciousness and/or sentience

are not required for transformative AI to be a concern.2. Separating the issue of potential harms to a sentient AI (an “Artificial Moral Agent”) from the issue of potential harms to humans as a result of transformative AI.While transformative AI is unlikely to happen in the next ten to twenty years, we should not rule out this possibility: there has been a recent acceleration in the rate of progress in AI, with many of the algorithms developed being applicable across domains. Moreover, technological breakthroughs are hard to predict. On September 11, 1933, Lord Rutherford, a Nobel prize winning physicist, dismissed the prospect of atomic energy as “moonshine”. Less than a day later, Leo Szilard invented the nuclear chain reaction.For these reasons, we dispute the claim “The probability of occurrence of such scenarios may from today’s perspective be very low.” It may be that with today’s technology the probability is very low, but similarly from today’s perspective the probability of catastrophic climate change is very low, because as of today it hasn’t reached catastrophic levels. Taking into account further development and future technologies (analogous to climate forecasts), we believe it is extremely likely that transformative AI will eventually be developed. There is a huge commercial incentive to do so, slow but steady research progress has been made, and there is no sign of any insurmountable obstacles. Given this is both the goal of many AI researchers and the likely long-term direction of the field, we believe it is worth considering this prospect now, and hope the EU can be a thought-leader in the ethical development of transformative AI.Feedback on other parts of the document:We would also like to highlight a few particular statements in other parts of the document that we believe should be revised or clarified:- In the Executive Summary, it is stated that “on the whole, AI’s benefits outweigh its risks”. Since the outcome of AI will heavily depend on policy decisions, we don’t believe it is possible to say whether the benefits will outweigh the risks before policy options have been thoroughly explored and analyzed. As an analogy, it’s possible that many European governments in the 1950s would have said that nuclear power’s benefits outweigh its risks; but facts seem to contradict that assessment: nuclear plant construction has essentially collapsed since Chernobyl, and more nuclear power reactors have closed than opened in recent years. Moreover, several major countries are completely opposed to nuclear power (Australia, Austria, Denmark, Greece, Ireland, Italy, Latvia, Liechtenstein, Luxembourg, Malaysia, Malta, New Zealand, Norway, Philippines, and Portugal have no nuclear power stations and remain opposed to nuclear power. Belgium, Germany, Spain and Switzerland have announced plans to phase out nuclear power completely.)- We are pleased to see the list of Requirements for Trustworthy AI, and in particular would like to emphasize the importance of Robustness and Safety.- In section I 3.2, it is stated that “freedom of the individual requires protection from direct or indirect coercion, surveillance, deception or manipulation”. We strongly agree, but would welcome further details about how this might be reconciled with freedom of expression, or how one distinguishes between persuasion and manipulation.In

section I 4, it is stated that “AI systems should not harm human beings”. There are many cases where one cannot tell in advance whether a given action - say, driving a human passenger to the airport - will result in harm. Suppose the directive were stated as “AI systems should not take any action that has some probability of harming human beings.” In that case, all autonomous vehicles would refuse to take human passengers or to drive on roads where humans are present. That would indeed be a reasonable outcome in cases where autonomous vehicles are considerably less safe than human drivers, but an unreasonable outcome in cases where they are considerably more safe. To have practical significance, we believe this statement needs to be reconsidered.- We would suggest rewording “in deviation of Fundamental Rights” (section I 5.3) with “in violation of Fundamental Rights”.In the glossary, AI is defined as “systems designed by humans that, given a complex goal...” This is a classical view of AI. Because of the problems with optimizing fixed objectives, which may not be aligned with true human preferences, we believe that other approaches (such as those that dynamically learn human values and preferences, rather than being given a goal) may be preferable. We recommend expanding the definition of AI to include these approaches. We would also recommend acknowledging that in the future, AI systems may be designed by other AI systems.- Section II 2.1 mentions “behaviour boundaries that must not be trespassed”. This assumes that humans are smart enough to write foolproof rules. Unfortunately, experience shows us that we often fail at this. As a mundane example: after 5000 years we still seem unable to do this in the area of tax law.- We would very much like to see the guidelines address the issue of impersonation and deception of humans by AI systems, and the implications of this for human dignity. We recommend consulting California’s Senate Bill No. 1001, which requires that bots disclose themselves when interacting with humans.It is important to note that we have already experienced some of the powerful effects of AI systems, namely the polarization and in some cases near-destruction of democratic societies by bandit and reinforcement learning algorithms operating in social media. We urge the AI HLEG to consider the extent to which these guidelines would have prevented or mitigated this outcome, and what can be learned from examples like this and those mentioned previously such as nuclear power.Thank you for giving us the opportunity to provide our feedback. As mentioned, we believe that the EU can be a thought-leader in the ethical development of transformative AI, and we offer our expertise to help shape these guidelines for trustworthy AI.Signed:Stuart Russell, Professor of Computer Science, University of California, Berkeley; co-author, Artificial Intelligence: A Modern Approach, the standard reference in AI; Founder and Director of CHAI; UK citizen; Chaire Blaise Pascal in Paris and French resident 2012-14; member of France’s AI International Scientific Board advising Pres. Macron.Mark Nitzberg, PhD, Executive Director of CHAIRosie Campbell, Assistant Director of CHAI; UK citizen; BA and MSc, University of Bristol.Joe Halpern, Joseph Ford Professor of

Computer Science, Cornell University; Fellow, American Association or Arts and Sciences; citizen of US and IsraelJuliana Schroeder, Assistant Professor of Management of Organizations and Psychology, Haas School of Business, University of California, BerkeleyAdam Gleave, PhD Researcher at CHAI; UK citizen; BA and MPhil, University of Cambridge.Rohin Shah, PhD Researcher at CHAI; BSc, University of California, Berkeley.Michael Dennis, PhD Researcher at CHAI; BSc, DePaul University, Chicago.Charis Thompson, RQIF Professor, London School of Economics and CHAI affiliate

Wouter	Termont	Centre for Logic and Philosophy of ScienceKU Leuven (University of Leuven)	Dear members of the HLEG AI,In the following text I would like to contribute my thoughts on your draft of the Ethics Guidelines for Trustworthy AI. My apologies in advance for posting my entire text in the general comment section; I wrote these observations before looking at the consultation form. Most of my commentary has to do with the document as a whole anyway. I will begin with some minor remarks and continue with some more substantial observations on the draft. Before I start, however, let me congratulate you on approach taken and the structure of the work already done, and thank you for the opportunity to help shape the final version of the document.While I expect some of these remarks to already be pointed out by multiple other commentators, I just mention some minor issues to start with. The first is a more general observation, pertaining to some of the distinctions made in the text, especially the different requirements for trustworthy AI and the (non-technical) methods of assessment. The rationale for their number and distinction is not always explicitly made clear, with the effect that some of them seem to overlap. As an example: a lot of the requirements seem like they should be part of robustness, and multiple methods seem to be about the relation between internal and external standards/panels/institutions. A more delineated distinction would sometimes help for clarity.Secondly, while the introduction to rights, principles and values characterizes values as concrete workable derivatives of principles, the 9 principles of the EGE are "based on the fundamental values laid down in the EU Treaties and in the EU Charter of Fundamental Rights", which seems to take the other way around. I am not sure if this is a mistake, but it does cause some confusion given the effort of describing their (ordered) relation earlier in the text.Another possible mistake is situated in the description of human-centric AI in the glossary: it reads "AI should not be seen as a means in itself." This should probably be "[...] as an end in itself." Similar sentences can be found in at least one other place (the executive summary) in the guidelines.Further, there seem to be some differences and inconsistencies between the definitions of AI in different documents, as well as within the HLEG Definition of AI document (mostly because of the update at the end). While defining the subject concept is definitely a necessity, giving different or inconsistent definitions clearly defeats the purpose. Since this issue does not pertain to the Ethics Guidelines, and it is quite complex, I will not go deeper into it. There exists numerous decent definitions within the extant literature on AI. That also brings me to a next point: the lack of references to used sources. Since I expect the HLEG to be informed by the many research on AI, I presume that a lot of what is written (and definitely the definition(s) mentioned above) should contain references. They would certainly have come in handy for commentators, to know on which material the draft was based.Another cause of confusion with definitions is the combined use of "explainability", "explicability", "traceability" and "auditability". Sure, in a footnote, it is explained that Floridi uses "explicability to refer both to "intelligibility" and to "explainability", which would capture the

need for transparency and for accountability. There is no difference here, however, with the much more current “explainability” itself, which also entails a kind of intelligibility, transparency and accountability. While I would therefore prefer the less exotic and more established term, the main point is to use it consistently throughout the document (and also in the definition document). Moreover, both “explainability” and “explicability” lie, in my opinion, too close to “traceability” and “auditability.” All of these terms stress the same aspect of transparency, namely that “AI systems should document both the decisions they make and the whole process that yielded the decisions [to] tell us how [a certain decision] came about,” i.e. “to be able to understand why it had a given behaviour and why it has provided a given interpretation." On the whole, the document maybe lays too much stress on this idea, stating things like “laypersons should be able to understand the causality of the algorithmic decision-making process.” Most people wouldn’t even understand how their pocket calculator works. Moreover, the main reason why AI systems are called ‘autonomous’ is precisely because they run processes so complex that we can’t follow them. Real “traceability” is therefore either impossible or useless to us. “Explainability” in a more free sense (like the way humans can “explain” their actions), on the other hand, defeats the purpose, since such explanations are often just-so-stories. In that light, being “demonstrably worthy of trust” seems like a contradiction in terms.To conclude the minor remarks, the term "AI ethics" seems somewhat ill-chosen, since it is ambiguous between the ethics present in the behaviour of AI systems, and the broader (human) study of ethics pertaining to AI. The latter, which is what the draft document is about, is a little less ambiguously identified with the term “ethics of AI.”Turning towards some more substantive observations about the draft, I would like to start by pointing out the level of generality on which the text remains. A lot of the requirements for trustworthy AI (accountability, non-discrimination, respect for privacy, robustness, safety, transparency, design for all, resilience to attack …) should be met by any product or service. Granted, some of these issues are especially pressing for AI technology, but in the current state they seem to take up the central focus of the draft. Such guidelines might make the text read as a disinterested “nothing new.” Shifting the focus to more AI-specific issues might be necessary to keep the guidelines a useful document for developers and deployers.As already touched upon above, one of the main AI-specific characteristics is the aspect of autonomy. The definition document specifies AI as “systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals,” and “[systems] deciding the best action(s) to take (according to pre-defined parameters).” The document also mentions “the ability to choose the best action to take in order to achieve a certain goal,” or more elaborate: “perceiving the environment [...], reasoning on what is perceived, deciding what the best action is, and then acting accordingly.” These sentences are, in my opinion, contradictory. Either an action follows some

autonomous decision or choice, or it is the effect of behaviour according to some pre-defined parameters. All current AI technologies fall in the latter category. This fact is alluded to in the document: "a decision, content, advice, or outcome, is the result of an algorithmic decision of any kind." However, if these "decisions" are nothing more than complex algorithmic processes, which (given the same input) are deterministic up to (pseudo-)randomness, then they are not decisions at all, since a decision can only be made with real autonomy. In their Statement on Artificial Intelligence, Robotics, and 'Autonomous' Systems, the European Group on Ethics (EGE) in Science and New Technologies correctly state that (current) AI systems cannot have (real) autonomy. They characterize "autonomy" as a philosophical term linked to certain cognitive capacities like consciousness and self-authorship, that are typically reserved for human beings. They claim that "it is therefore somewhat of a misnomer to apply the term 'autonomy' to [...] very advanced complex adaptive or even 'intelligent' systems." In contrast, they say, the literature and debate on AI uses this term to "refer to the highest degree of automation and the highest degree of independence from human beings in terms of operational and decisional 'autonomy'," which is the notion of autonomy also referred to in footnote 24 of the draft guidelines. In this latter sense, AI systems are 'autonomous' as far as they do not rely on human oversight or control. The problem with this confounded use of the term 'autonomous' is twofold. First, it has a dubious effect on the closely related concepts of responsibility and (legal) accountability. This effect is already present in the draft guidelines, for example in the assessment method for trustworthy AI by governing AI autonomy: it talks about "measures [...] to ensure that an AI system always makes decisions that are under the overall responsibility of human beings." With contemporary systems and technologies, responsibility and accountability, there is no other possibility, which renders the method vacuous. It is not the case that human autonomy is in any way reallocated to AI systems. Contrary to the way the terms are used in the draft guidelines, autonomy is not equivalent to agency: the latter can be delegated to AI systems, so that they do stuff that we no longer (have to) do. Equating the Principle of Autonomy with "Preserve Human Agency" is therefore wrong. While the division of responsibility and accountability among multiple human beings or human aggregates (groups, institutes, companies …) indeed becomes more interesting because of the complexity and power of AI, it is of exactly the same nature as division of responsibility and accountability in any other domain. The only relevant cases are, after all, those where robustness fails or, as the draft states, when "[t]hose in control of algorithms [...] intentionally try to achieve unfair outcomes." Second, the EGE goes wrong, in my opinion, with the claim that there will never be a system that can be called 'autonomous' in the original (philosophical) sense. Such claims are mere conjectures based on intuitive modal thinking. The issue I have with this narrow idea of autonomy is most visible in the following citation: "Since no

smart artefact or system - however advanced and sophisticated - can in and by itself be called ‘autonomous’ in the original ethical sense, they cannot be accorded the moral standing of the human person and inherit human dignity.” It is perfectly understandable, given the current state of the art, that a narrow vision of autonomy is the most intuitive; after all, at the moment we only have “Narrow AI.” However, while the HLEG definition document still mentions this distinction, the draft guidelines lack any consideration for the possibility of “Strong AI.” This neglect is, in my opinion, extremely dangerous, especially in guidelines on the level of the EU, since it makes us disregard any development that goes beyond the foreseeable extension of current technologies. In particular, it makes the guidelines useless in case the genesis of “Strong AI” occurs without regard for the charter.Furthermore, a human-centric approach to autonomy goes against the European values and principles the charter claims to be based on. While the EU Treaties and the Charter of Fundamental Rights indeed focus on its human citizen’s well-being, human primacy was never the origin of these values. The modern principles that lie at the birth of the EU are enlightened ideas of inclusiveness and consideration, not necessarily limited to human beings. The one place where this true nature shimmers through the text is in the Principle of Non-maleficence, where concern for harm to animals (and the environment) is expressed. Speaking of the “potential harm associated with [...] the development of Artificial Consciousness”, without mentioning the other side of the coin, thus disregards these central European values. From footnote 18, it is clear that the HLEG know this. It makes me wonder why, in a document full of reference to rights, principles and values, this important aspect of AI is passed on so lightly. Claiming that “[strong AI] might lose alignment with human values,” (footnote 20) or that “[t]he development [of such systems] would potentially present a conflict with maintaining responsibility and accountability in the hands of humans, and would potentially threaten the values of autonomy and self-determination,” (footnote 19) only makes it worse. Indeed, it would threaten these values, but not at all because it conflicts with human-only responsibility. Rather, the values would be in danger because we would fail to grant responsibility to systems that are really autonomous. The fact that “[w]e currently lack a widely accepted theory of consciousness” makes it all the more pressing to carefully consider this possibility. More research into consciousness and its connection to moral standing is therefore a necessity. The conclusion of the draft reads: “Europe has a unique vantage point based on its focus on placing the citizen at the heart of its endeavours. [...] This document forms part of a vision [...] which will enable Europe to become a globally leading innovator in AI, rooted in ethical purpose.” Given the points I put forward above, this conclusion can no longer be true. There are other guidelines already in existence that go way further in their ethical inclusiveness (for example, the South Korean Robot Charter). If we really want to maintain our human dignity, we need to start approach AI in a different way than we have done towards other entities

(human minorities included) throughout history.Real trust is built through interaction. In order to achieve it, AI systems will have to act morally correct. Real trust, however, is also reciprocal. If we ever create real autonomous AI, we will only achieve trust if those systems can trust human beings to act equally moral towards them. A framework which seems especially suited in that aspect is virtue ethics, circumventing the infeasibility of normative design in deontological approaches or utility design in consequentialism. While mostly referred to as an ethical framework for the human handling of AI technology, I believe it to be a crucial approach to ethics of AI with the possibility of real autonomous AI in our minds. Ethical values need to be learned through (reflection on) experience with real moral dilemmas. I therefore suggest that virtue ethics, or more general ethics through education, be added to the non-technical methods of achieving trustworthy AI, and that research into technical support for this be added to the technical methods.Note that the bulk of the above observations only apply to the second issue with using the term "autonomy," i.e. the guidelines should foresee the genesis of real autonomous AI. Since the current situation is different, however, the rest of the guidelines should be clear about the lack of any autonomy in current AI systems. The ethical concerns should then no longer just be focused on the aspects of responsibility and accountability, which leaves room for one last ethical topic that is somewhat lacking in the draft. A lot of stress is put on the goal "to identify how AI can advance or raise concerns to the good life of individuals" and the connected claim that one of the pillars underlying the Commission's vision for AI is "preparing for socio-economic changes." From chapter 1: "Ethical insights [give us] finer grained guidance on what we should do with technology for the common good rather than what we (currently) can do with technology," and "AI systems should be designed and developed to improve individual and collective wellbeing." However, none of these aims are really worked out in the guidelines, which can be characterized defensive rather than advancing, conservative than progressive. Nowhere does the document define any socio-economic goals the EU might want to look into herself.The single attempt at formulating how AI could be beneficial on a European scale, is "AI systems can do so by generating prosperity, value creation and wealth maximization and sustainability." This statement neglects everything we know about the effect of AI technologies on the job market. "Wealth maximization" does not belong in that sentence. If one thing should be clear in the prospect of AI, it is that we need to drastically change our economic model. AI can be a powerful technology, bringing profit and high GDP's, but if those are the measures of Europe's value and prosperity, they will also be the measures of our inequality. On the other hand, AI can be the perfect tool for reinventing our idea of "wealth" and rethinking our social structure. The ethical guidelines seem like perfect place to start moving towards that dream. To summarize my comments on the HLEG Draft Ethics Guidelines for Trustworthy AI: I started with some concrete minor remarks and pointed out some more substantial

								issues with the text. The former included the vagueness of some distinctions (e.g. the requirements and methods of assessment), an inconsistency in the use of “rights, principles and values”, a possible mistake between “means” and “end”, the inconsistencies between definitions of AI, the lack of references, the confusing use of terms related to “explainability”, and the ambiguity of the term "AI ethics".As more substantial contribution, I first criticized the level of generality of the text, and suggested a shift towards more AI-specific issues to keep the document relevant. In this light, I pointed out the ambiguous and even contradictory use of the AI-specific term “autonomous”. I argued that since all current systems lack real autonomy, we should shun such language, and realize that the issues of responsibility and accountability are simply about their (complex) division between human beings. On the other hand, although no current systems have real autonomy, I argued that sufficient attention should be paid to the possibility that in the future some systems will. I based this last point on two reasons. The first is simply the danger of limiting the applicability of the guidelines to currently intuitive cases. The second reason is that the European values and principles, on which the guidelines are based, call for a more inclusive and considerate approach. I suggested virtue ethics as a possible approach, but more research into autonomous AI is definitely needed. Finally, in my last observation I shortly advocated a stronger focus on possible socio-economic goals the EU reach with AI, since simply aiming for ”wealth maximization” is obviously not enough.I hope to have suggested some useful practical changes, as well as some possible general issues with the draft, and look forward to read the final version.Sincerely,Wouter TermontCentre for Logic and Philosophy of ScienceKU Leuven (University of Leuven)
Lotta	Engdahl	Ministry of Transport and Communications	The Comission’s High-Level Expert Group on Artificial Intelligence (AI HLEG) published in December 18th 2018 a draft of the AI Ethics guidelines for Trustworthy AI. The final version is due in March 2019. We welcome the High-level group’s work as guidance on the concrete implementation and operationalisation into AI systems presented in the working document. The document is a starting point on guidelines, that should be regularly updated over time to ensure continuous relevance as the technology and our knowledge thereof evolves. The document should therefore be considered as the first step for the discussion on “Trustworthy AI made in Europe”.Artificial intelligence (AI) are largely based on the efficient use of data. The development of the data economy increases growth, employment, innovation and the human-centred focus of the global data economy. The goal must be to build a competitive and human-driven data economy in the EU which must be based on extensive availability and use, while respecting the rights and privacy of individuals. The practices of human-centred data management (MyData) shall be strengthened so that individuals have the ability to share and enable efficient use of data for better and individual services. Business should be enabled by creating	We share the view that AI holds the promise to increase human wellbeing and the common good, but to do this it needs to be human-centric and respectful of fundamental rights. A digital society must be built on solid trust. When data is used as a raw material for services, it is important to increase trust among users towards products and services. Realising trustworthy AI means individuals right to transfer their data on one service to another in a usable format. Therefore human-centred data economy should be strengthened by creating concrete mechanisms for interoperable structures for data transfer.According to our view, the five ethical principles in chapter I on ethical principles in the context of AI and correlating values (do good, do no harm, preserve human agency, be fair and operate transparently) mentioned, is supported. For citizens the most valuable value is ‘do not spy’, which should be added in the list of values. Strong emphasis should be put on transparency. For example, transparency of the algorithms used by automated vehicles should be ensured. This would provide authorities full understanding and make it possible for them to influence the criteria on which the decision-making of the vehicles’ automated driving system is based, as well as to increase trust amongst road users.	We see the requirements of trustworthy AI listed in the chapter II as a good start to realise the impact of AI in the future. These requirements are:1. Accountability 2. Data Governance 3. Design for all 4. Governance of AI Autonomy (Human oversight) 5. Non-Discrimination 6. Respect for (& Enhancement of) Human Autonomy 7. Respect for Privacy 8. Robustness 9. Safety 10. Transparency The listed requirements are supported and a concrete start when developing, deploying and using AI. Nevertheless the list should also include access to data as one of the requirements. Access to data is the key enabler of AI. Market development should be guided towards decentralized data solutions where data is managed by data access and identity management to be used in digital service networks. Data should be opened on fair, reasonable and non-discriminatory terms. True personalization of services can be achieved by building data transparency and data flows via decentralized data solutions (like blockchain) that connect data economy systems ensuring interoperability and usability of data while preserving data protection and privacy. Blockchain offers also tracybility in terms of to the automated decision-making. This should be adressed thorougfully in the chapter II Non technical	Autonomous driving/movingTo select autonomous driving/moving as a use case of AI in the draft of this ethics guidelines is very much encouraged in the path of deploying the AI system in the future. We strongly recommend to use the term Autonomous Transport instead of Autonomous driving/moving. Autonomous transport is one of the first fields where AI solutions will be implemented in real life. For example shuttle buses and high lane buses are early adapted means. Artificial intelligence is a key to ensuring the ambitious goals to continue to advance sustainable and safe autonomous traffic in different modes of transport. Data generated for and from the use of automated vehicles as well as collected from traffic, especially by vehicle manufacturers, should be made accessible for the use of third parties and authorities in order to ensure the fluidity, compatibility and safety of automated traffic. As already mentioned in chapter II 1. requirements of trustworthy AI transparency of the algorithms used by automated vehicles should be ensured in order to provide authorities with full understanding and make it possible for them to influence the criteria on which the decision-making of the vehicles’ automated driving system is based, as well as to increase trust among	For further discussion, there is a need for frameworks to identify the different levels and to bring up the essential issues related to development of AI. It is important to evaluate the development of AI on a general level and sector specific-level paralelly. The development should be considered together to minimize the gap between regulations on these levels.The level of ambition should be raised by building a sustainable, competitive and human-driven data economy in the EU, which must be based on extensive availability and use of data, while respecting the rights and privacy of individuals. All new initiatives should also take into account the horizontal interoperability and free flow of data in the single market. Business should be enabled by creating common principles, practices and rules for data sharing and access. If these lighter means are not enough, development can be promoted through regulation.Promoting AI and automation is one of the central means of improving the competitiveness of the EU as part of the global data economy. Technological innovations, such as the Internet of Things and AI, enable the collection and use of ever-growing data volumes and the building of key societal services (e.g. smart traffic, healthcare, trade and insurance) on top of this data. We need

			shared principles, practices and rules for sharing data and granting access rights. This should concretely be delivered through the process. The cooperation on AI intelligence is not only important on EU level, but also important on global level. Europe has to seize the opportunity for trust-based business development. This is also our strength to stand out from other global players in AI. In data economy we need to move our focus from business of scale to business of skills. Consumers rely on big companies and businesses. Therefore we should enforce and strengthen the possibilities for small business to enter the market. Individual trust can be strengthened by adapting shared requirements, principles and understanding of the functionalities of businesses.Transparency serves as a basis for the reliability of an artificial intelligences decision-making. Reliability can be achieved by enabling an external entity to review and evaluate the basis for descision-making. Also, transparency of the algorithms used by AI should be ensured. Only then we can provide authorities with full understanding and make it possible for them to influence the criteria on which the decision-making system is based on, as well as to increase trust amongst individuals. Global markets makes sure that automation is one of the central means used in future global data economy as AI systems uses automated data to achieve the given goal or action. Automation also helps to improve the safety and efficiency of transport and the fulfilment of environmental and climate goals in all modes of transport. Therefore AI also has a huge meaning in a broad societly context.Automation also applies to the field of humanities, because if you want to strengthen citizens' trust, then action must also have acceptable and solid reasoning. Individuals have poor prospects for anticipating future development and therefore a change in the operating environment requires extensive research and updating of skills. Human behavior also changes in a new situation. In addition to researching artificial intelligence, ethical issues should also be systematically in focus of research and studies.	methods. It is important that the data gathered is available for use by other services and service providers. For example in transport market the requirements on access to data about routes, parking, scedules and pricing, should be available for other service providers. Also vehicle data should be available for neccesity extent. Therefore interoperability between for example the transport services should be opened. Access to such information should be provided through an interface in the data system. The storing and sharing data should be based on so called on-stop shop principle, i.e. storage of data in one system only should be made possible in order to enable access rights to parties wich have the right to access such data. Infrastructuring a digital twin from data gathered by vehicles (for example weather and road conditions) should be available and opened. The opening of interfaces should be done in a technologically neutral manner.	road users. According to our view, the ten requirements of Trustworthy AI are all important to support the operationalization of a further AI. We would like to address a specific glance on the means to prevent faults and wrongly made decisions by AI, instead of paying to much focus on the accountability and monetary compensation mechanisms. When it comes to complex AI systems such as autonomous driving, auditability of the AI system becomes critical and it is essential to have pre-explanationed the methods, logic and the decisions it makes. When the functioning of AI systems can be evaluated against a certain set of criteria and simulated ex ante, then its proper functioning can also be evaluated ex post.	to strengthen the trust of consumers in technologies, devices and applications related to the safe processing of data. In the societal context transparency of algorithms is necessary to enable discussion about the influence of the decisions and effects of machines on people's lives and thereafter if necessary regulation is in order (e.g. traffic safety of autonomous transport).
Johannes	Hofmeister	Chairman of the T-Mobile Austria Works Council	- We welcome the possibility to contribute to the stakeholders' consultation and underlines the importance of a broad public debate and information on AI. This debate must result in clear ethical and social guidelines and standards with the aim of improving the living and working conditions of European citizens. - We acknowledge the innovative potential of AI and new technologies that can be beneficial for our society. However, these new technologies also create challenges and we are concerned about the possible risks and consequences relating to working conditions, skills and training, ethics, equality, health and safety (among others). Therefore, we would like to underline the importance of addressing AI technologies and robotization as topics for collective bargaining at all levels (company, national and European). AI and robotics have a huge impact on the future labour market, as jobs will sometimes disappear or be transformed and other jobs will be created. We need to accompany this process and address the	- we support the human-centric or human-in-command approach suggested in the guidelines. We agree that it is necessary that humans always need to remain in control of technology and machines. Likewise, we agree that the use of AI needs to respect European values and fundamental rights. - We recommend an expansion of 3.2 "In an AI context, freedom of the individual requires protection from direct or indirect coercion, surveillance, deception or manipulation." Much profiling that lies at the heart of AI systems relies on a degree of manipulation of data. This is not least relevant in relation to the world of work and especially in the use of AI in Human Resources. - We welcome that the HLEG understands the need to ensure that those involved in the development and marketing of AI (researchers, engineers, designers etc.) act in accordance with ethic and social responsibility criteria. This should be addressed by changing educational priorities for technical subjects and by providing	- In order to achieve "trustworthy AI", we need to establish public, independent and autonomous organisations that can control and audit (labour) algorithms (e.g. to identify underlying biases and the objectivity of data sets that train algorithms). Likewise, the implementation of the ethical guidelines on AI must be monitored. A European observatory focusing on the ethics in AI systems could play the role of an independent watchdog, including in business. - We would like the advice „to always keep record of the data that is fed to the AI systems" from the heading of data governance included under Accountability. For workers, it is paramount that the datasets used to evaluate performance, or in hiring or firing processes is transparent and can be accounted for. - The explanation of the principle of autonomy covers the question of AI at work only in a footnote, whereas this is an important issue that should be given a more prominent place. We would like to highlight	- We welcome that processes shall be examined in order "to allow a human control, if needed" (assessment list – governing AI autonomy). In so doing it should not be a question "to keep a human in the loop". We need clearly defined measures, that empower people to exert this control in all processes – regarding resources (technical equipment etc), organisational needs (time, liability, etc.) and qualification. - We suggest extending the list on the assessment of use cases (p.28) and add the question of processes, in order to use AI to ensure decent work (development and impact assessment).	- We welcome the call for Accountability Governance on page 21. The establishment of Data/AI Governance Councils in companies will indeed strengthen the accountability of AI systems and will address a weakness in the GDPR. The Council should consist of shop stewards and management and be responsible for holding management accountable and transparent to the use of AI and data. Whistleblowers should be able to address concerns to the Council and mandate the council to investigate on reported issues. - We welcome the process of developing guidelines for a trustworthy AI made in Europe, which encompasses corresponding "guidelines made in Europe", but would like to raise the question why non-European companies such as Google were granted full membership and full participatory right in the High- Level Expert Group. The status of associate expert would be more appropriate.

question of skills and training for the future workforce: need to ensure that training on necessary digital skills is provided by education institutions and companies, and that it is not the sole responsibility of the worker to keep up with the rapid technological developments. Employability needs to be promoted through upskilling and reskilling schemes for workers. Investment in formal, informal and life-long learning is key; we must enable people to work with AI or invest in competences that AI will not cover. It is important to develop action plans at EU and national level together with education providers and social partners in order to modernize education and vocational training. We therefore welcome the call from the ILO Global Commission on the Future of Work for "a formal recognition of a universal entitlement to lifelong learning and the establishment of an effective lifelong learning system. (https://www.ilo.org/global/topics/future-of-work/publications/WCMS_662410/lang--en/index.htm)
- The social partners play a key role in this and the EU should cooperate with them and national governments in order to identify which job sectors will be affected by AI. We need to understand the timeline and extent of changes in the labour market. The involvement of social partners is a must to find appropriate and future-proof solutions to concerns relating to employment, training, the nature of work, (in)equality or social systems and collective bargaining, especially at sectoral level.
- As AI and automation have the potential to transform not only simple tasks but very complex processes, we need to have a large public discussion about the areas in which the use of AI is reasonable and beneficial for society. Part of the debate should be the question of how the profits generated by AI should be re-invested for the common good by creating employment in domains such as care, health services, education or mobility. Employees should participate in the distribution of profits, e.g. through wage increases or reduction of working time. Moreover, AI wins should be used to strengthen social security systems. This could be a measure to address the problem of future job losses and the precarisation of employment relations in a-typical work (e.g. platform work) due to AI and automation.
- It is therefore important to integrate the aspect of the quality of jobs, decent work and social progress into the ethical approach in order to create a balance with the purely economic objectives underlying the creation and use of AI and robotics.
- The Human-centric approach (HCD) not only presupposes information, transparency, participation and traceability, but also requires specific negotiation processes regarding decision-making in view of the aims and implementation of AI-systems at a very early stage for stakeholders such as employees and their codetermination bodies. lifelong learning opportunities (e.g. by incorporating ethics and the humanities into training courses in engineering).
- Organisations and companies should develop tools to facilitate ethical discussions and decision-making throughout the whole design process. This should be completed by internal training programs on ethics for all employees. Such training should help employees understand the AI systems themselves, their rights in relation to said systems and their possibilities of redress, complaint etc.
- AI should provide an opportunity for workers to apply their skills and competences to the fullest while at the same time remain owner of the production process. This includes the principle of transparency in the use of AI systems in HR, like the hiring of employees or the performance assessment of staff. It is important to safeguard the rights and freedoms of employees in line with non-discrimination rules as regards the processing of workers' data.
- we welcome 5.1 – 5.4. We support that these examples raise real-life concerns of the adverse consequences of AI systems.
- In 5.2. we urges the group to expand on the issue of the human's right to know they are interacting with an AI identify. This could be done through a "labelling" system. For example, online bots should be labelled as such. Users should be made aware of the use of bots and AI in customer call-centre or help desks etc.
- We would welcome that the employer-employee, employer-worker relation is explicitly mentioned in 5.3 as an example of power asymmetry.
- Taking into account the power asymmetry in employer-worker relations, a separate point 3.6 on "workers' rights" should be added, which should contain the following points: "decent work by design", equal negotiation processes in the sense codetermination rights, informational self-determination of employees, non-discrimination principle and freedom of association including the right to strike. This is needed in order to secure worker's rights to co-decide on aims and application of AI systems, and create a legal framework.
- Concerning the long-term risks and concerns we welcome that these should be considered. This could become an integral part of the accountability and audibility demands – i.e. that developers, users deployers etc need to reflect on the development/changing nature of the adopted AI as well as engage in predictions/forecasts of its future development scope and the consequences (positive and negative) hereof).
- AI's influence does not only affect the world of work, but also democracy and society as a whole. We welcome that the draft refers to this point in Chapter I, paragraph 5.3., by stating that AI is not to be implemented in order to enable "citizen scoring" by a state/government. But this should also apply to private businesses. Neither states nor companies should be allowed nor have the possibility, to create human profiles such as "moral personality" or "ethical integrity". We reject the proposed opt-out-function and even possible "opt-in"-functions are not to be designed in a way that they conflict with fundamental human rights and possibly lead to the waiving of the right of workers to individually and collectively opt out or withdraw from the use of AI systems (or a decision chosen by an AI system) if they undermine the workers' autonomy, decision making competence or disrespect fundamental rights and ethical principles. We recommend the inclusion of a special chapter that provides for ethical guidelines on AI in the work environment to address these issues more in detail.
- we welcome that the HLEG on AI acknowledges the importance of social dialogue to realise trustworthy AI. We would like to add that the involvement of social partners, and in particular employee representatives, should not only take place regarding the general public debate on AI. Social partners should be involved in the establishment of codes of conducts, of standardisation schemes, development of training and in the proposed accountability governance. Employee participation and inclusion should take place early in the design, development and deployment of new technologies including AI and robotics. It is essential and important not only to inform and consult workers representatives in the work place or at branch level, but to enhance their co-determination rights and ensure their right to co-decide on the aims, reasons and implementation of AI at the workplace.
- Social partners at all levels should be involved in the implementation at company, industry, national or international level, including through collective agreements setting standards. In this context, it is required to describe the negotiation processes, e.g. central control structures for sector-specific solutions (cf. ‚AI Now' Report 2018): „Governments need to regulate AI by expanding the powers of sector-specific agencies to oversee, audit, and monitor these technologies by domain." (https://ainowinstitute.org/AI_Now_2018_Report.pdf)
- Regarding the principles of accountability and transparency, we need to establish mechanisms for the protection of whistle-blowers who disclose the risks of AI systems or the non-respect of ethical principles – especially in the case of employees in companies that develop such systems. Internal reporting of risks and violations should be supported and rules in place to ensure follow up.
- Organisations and companies should pay attention to potential biases encoded in the system development, training data and model performance – especially those that my affect the most vulnerable. They could also establish an internal ethical review process to democratise the decision-making process
- Companies should not only increase transparency regarding the design and development of AI systems, but also in organisational chains of responsibility.

			services that are useful for a person. AI-based services, that are important for work and life, must be designed in such a way that they do not require the collection of data which could be useable for human profiling. - Creating big data-bases always includes the risk of hackability as well as intentional and unintentional data-leaks. The guiding principle of "data-sovereignty" needs data-security in order to be viable. This implies explicitly not surveying data in areas that are of highly explosive nature for people in e.g. political, private or work-related areas. Fundamental rights as informational self-determination, the freedom of association and freedom of speech are not to be put at risk by creating such data-bases.				
Balint	Toaso	Joint comments by dr. Bálint Tóásó, Levente Szabados, dr. Ákos Krénusz, dr. Fanni Márkus, dr. Gergő Szalai-Bordás	Dear Ms. Smuha,with reference to the Draft Ethics Guidelines for Trustworthy AI prepared by the European Union's ("EU") High-Level Expert Group on Artificial Intelligence ("HLEG") dated Brussels, 18 December 2018 (the "Draft"), the team of the above authors (jointly "we"), would like to take part in the consultation and share our views around regulatory and ethical aspects of artificial intelligence ("AI") and its future deployment in the EU.As per the request of HLEG, we will provide our comments to the Draft broken down to chapters of the same. However, as Chapter II and III of the Draft are built up around the same 10 requirements, for the sake of readability, we aggregated our remarks to those and summarized them under one section. Accordingly, our comments are structured as follows:SECTION NO NAME OF THE RELEVANT PARTSection 1 Glossary, Rationale and Foresight of the GuidelinesSection 2 Respecting Fundamental Rights, Principles and Values - Ethical PurposeSection 3 Realising Trustworthy AI, Assessing Trustworthy AISection 4 General comments to the DraftSECTION 1: Comments to the Glossary, Rationale and Foresight of the GuidelinesWhile we understand that a separate document is currently drawn up about the definition of AI, we are of the view that we should still provide some details about its very characteristics. Therefore, we suggest complementing the definition part of the Draft as follows:Artificial Intelligence or AI: AI refers to systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data (by recognizing correlations and informative patterns), processing the data into informative representations, and using these to derive predictions about unobserved (unlabelled) or future data. This enables AI systems to reasoning on the knowledge derived from this data and potentially decide the best action(s) to take (according to pre-defined parameters) to achieve the given goal. Though complex actions can be integral part of AI systems, the core functionality of pattern extraction, representation and prediction is their defining characteristic.AI systems can also be designed to learn to adapt their behaviour by analysing how the environment is affected by their previous actions. As a scientific discipline, AI includes	SECTION 2: Comments to Chapter IRights:We believe that in order to ensure appropriate and adequate level of protection for EU citizens, we must understand the working mechanism of AI. Therefore, we suggest dividing the protective measures, requirements into two sets of requirements. The first set of requirements must ensure that any data collected for the purposes of development and functioning of AI, is collected lawfully and in compliance with applicable laws ("Knowledge Discovery Stage"). The second set of requirements should deal with the core functioning of AI, liability issues and attributability of failures, errors generated by AI, as a predictive model ("Functioning Stage"). We have summarized the two groups in Figure 1 below.Accordingly, we believe that the five families of rights identified by the Draft should be re-grouped in line with the very functioning of AI. A clear separation would be useful in order to create a more effective regulatory framework in the future. In general, we agree that each of the families of rights enlisted are fundaments of the European way of life.We understand that (i) human dignity, (ii) freedom of the individuals, as well as (iii) citizen rights seem to apply and reflect more to the Knowledge Discovery Stage. Thus, these must be developed with focus on the lawfulness of data collection, access to data, consents, the qualification of the internet and the digital space, being either private or public area, information obligations of those conducting surveillance towards data subjects, and the right of individuals 'not being subject to surveillance'. It should also be decided how we would like to proceed with surveillance that is likely not to result in the collection of personal data.Also, even if the data subjects are opting in to be under surveillance, prompt, appropriate and fair compensation for information vested in the data generated through the surveillance of the same must be ensured (either in form of digital tax, or compensation directly to the data subject).Whereas, the families of rights such as (i) respect of democracy, equality, non-discrimination, and (ii) solidarity, including the right of persons belonging to minorities should gain more ground during the Functioning Stage. These seem to be more relevant when setting the patterns and rules for the predictive structures, modelling, identifying the desired outputs and weighing the different characteristics and qualities in a	SECTION 3: Comments to Chapters II and IIIAs we have detailed above, we are providing our comments jointly Chapters II and III of the Draft.Accountability, Safety:First of all, we agree with HLEG on the importance of accountability and safety in terms of AI development and operation. To strengthen the public trust in AI solutions, EU's regulatory environment must have clear answers to questions arising in connection with AI, especially in case of its malfunction and/or if it causes damage.Therefore, it is indeed of high importance to identify a sound and stable liability regime which could successfully react to and settle the tort cases stemming from AI usage. Though challenging it may sound, we are of the view that this could be done while relying on already established schemes and structures. Most importantly, the liability regime applicable to a given AI solution should not be selected by the mere fact that AI is involved, rather by the risk and exposure of the very functioning of the given AI and the application territory instead. Why would we use the same liability regime for AI driven autonomous vehicles and marketing activities? By the risk and exposure towards the society triggered by the given application area, we should clearly differentiate amongst the applicable liability regimes.If the given type of activity requires so, for instance for automotive, continental legal systems usually employ a stricter liability regime (besides general civil law liability regime), which is applicable to the operators of certain dangerous activities. This liability regime is almost objective, making the exemption from the liability almost impossible. This view entails a proportionate and layered system of responsibilities depending on the sensitivity of the matters / potential for damage. Different legal use cases of AI should be assigned to set categories of responsibility that should also require given tools / models of transparency and explainability (see our remarks on model interpretability above).Distinguished attention should be paid to the dissenting features of AI, i.e. what sets it apart from other state of the art software solutions. We understand that this dissenting feature is its feedback loop, based on which AI machine learning is able to further develop its predictive model sensitivity. Regulators should first investigate the functioning of such feedback loop mechanism and identify human control and cross checking points and	SECTION 3: Comments to Chapters II and IIIAs we have detailed above, we are providing our comments jointly Chapters II and III of the Draft.Accountability, Safety:First of all, we agree with HLEG on the importance of accountability and safety in terms of AI development and operation. To strengthen the public trust in AI solutions, EU's regulatory environment must have clear answers to questions arising in connection with AI, especially in case of its malfunction and/or if it causes damage.Therefore, it is indeed of high importance to identify a sound and stable liability regime which could successfully react to and settle the tort cases stemming from AI usage. Though challenging it may sound, we are of the view that this could be done while relying on already established schemes and structures. Most importantly, the liability regime applicable to a given AI solution should not be selected by the mere fact that AI is involved, rather by the risk and exposure of the very functioning of the given AI and the application territory instead. Why would we use the same liability regime for AI driven autonomous vehicles and marketing activities? By the risk and exposure towards the society triggered by the given application area, we should clearly differentiate amongst the applicable liability regimes.If the given type of activity requires so, for instance for automotive, continental legal systems usually employ a stricter liability regime (besides general civil law liability regime), which is applicable to the operators of certain dangerous activities. This liability regime is almost objective, making the exemption from the liability almost impossible. This view entails a proportionate and layered system of responsibilities depending on the sensitivity of the matters / potential for damage. Different legal use cases of AI should be assigned to set categories of responsibility that should also require given tools / models of transparency and explainability (see our remarks on model interpretability above).Distinguished attention should be paid to the dissenting features of AI, i.e. what sets it apart from other state of the art software solutions. We understand that this dissenting feature is its feedback loop, based on which AI machine learning is able to further develop its predictive model sensitivity. Regulators should first investigate the functioning of such feedback loop mechanism and identify human control and cross checking points and	SECTION 4: General CommentsLet us also detail our general comments which we deem as applicable to the entire Draft.Pragmatist approach:Firstly, we completely agree that the following are the key pillars of the European AI ecosystem: (i) increasing public and private investments in AI to boost its uptake; (ii) preparing for socio-economic changes; and (iii) ensuring an appropriate ethical and legal framework to strengthen European values. In support of the first pillar, we are of the view that we should also consider the role and implications of AI from a global perspective. It should be underscored that there is a growing consensus that AI and the appropriate deployment and application of technology tools will heavily determine EU's transforming global role in the 21st century. There is a realistic threat that if Europe cannot unify its "AI forces" and jointly utilise its diverse and far-reaching know-how, as well as its funds backed by harmonised and strongly pro-AI regulatory landscape, we are going to be overtaken by many actors in the global AI rally. And this would heavily determine our ability to maintain prosperity. According to our expectation, AI is likely to expedite the pace of digital innovation and will be a key driver in EU's growth potential. As AI gains ground in more and more application territories, its significance increases with every day.We believe that traditional European values and our unique approach to human rights must be appropriately balanced out with EU's interest to remain relevant as a global player. We have to understand that AI itself, is not 'ethically or morally loaded', it is not good or bad in its very sense.Yet another digital and IT instrument that the humankind wants to use and deploy for its greater wellbeing. Also, AI, as such is a well-paying business instrument, just like cars, software or any other goods that are produced for sale. Therefore, we should try not over-emphasise its ethical characteristics. It should be just as ethical as already existing software solutions and human made decision-trees, knowing indeed its very characteristics and main dissenting features against already existing technological solutions.Regulatory considerations:We should not forget that wellbeing of European citizens is founded on economical pillars. And thus, we would need to adequately assess whether the overregulation of AI and consequently the narrowing down of application areas where

several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems).A separate document elaborating on the definition of AI that is used for the purpose of this working document is published in parallel to this Draft. Since the core value of such systems lie in their predictive abilities, the notion of "predictive systems" in general is to be emphasized, broadening the scope of this document so as to apply to other data or knowledge driven methods e.g., currently considered parts of "statistical methods". In this regard the ethical principles agreed upon have to apply to all "predictive systems" irrespective of their realization or current categorization as "intelligence".Also, for a more detailed introduction to the working mechanisms of AI, we suggest adding the following two definitions to the Glossary:Model Interpretability: In case of any complex model, questions of its trustworthiness rely on its interpretability. Different forms and levels of interpretability are feasible and desirable for certain application areas of predictive systems, depending on the needs for control and potential risks involved with the usage of the system. Though some form of interpretability has to be ensured in all cases of predictive models (so as to ensure trustability), this requirement should entail: (i) the strictest form of simulability, (ii) the less strict requirement of decomposability or Algorithmic Transparency, as well as (iii) forms of post-hoc interpretations, like explanation by example or description.Due to their general nature, our remarks to the rationale and to the foresight parts can be found under the general comments under Section 4 below.

given model. It remains to be answered what AI can do with anomalies and even errors in the patterns and historical data.In summary, we believe that it would be appropriate to apply such grouping to the rights of EU citizens, as it may better structure future discussions around regulatory aspects. Figure 1: Relevant stages of the functioning of AIEthical principles:The Draft enlisted five ethical principles and correlated values that must be observed to ensure that AI is developed in a human-centric manner. We believe that (i) the principle of non-maleficence: "do no harm", (ii) the principle of autonomy: preserve human agency, and (iii) the principle of explicability: "operate transparently" are the backbone of the future regulation. Whereas, there could be arguments that (i) the principle of beneficence: "do good" and (ii) the principle of justice: be fair, might be deemed excessive and be considered as overreaction of novelties tagged along by AI.AI, as such, is not per se morally or ethically loaded, it is rather a partially new type of digital, IT business. Also, it is a fair statement that it will be key factor in EU's competitiveness in the global rally. Following this line of thinking we could conclude that AI functions by business rationale, and therefore we cannot expect developers and companies to comply with unrealistic requirements. Even if it is still hard to assess the future role of AI in our life, AI, in our belief should not be fairer than any other lawful goods or services. It should not harm or breach any laws instead. If the later criteria are met, AI would qualify properly.Is car manufacturing fair, are the services of a hairdresser serving the greater good? We are of the opinion that if the "do not harm" and the "do not breach" criteria are met, AI should get the green-light, without directly serving any higher purpose. It must however be underscored that do not harm and do not breach principles must be adequately and accurately turned into clear and straight-forward normative rules backed by effective enforcement tools.In our assessment the principle of autonomy: preserve human agency, draws down to the question of liability and eventually to questions around obligatory AI liability insurance. Whereas, the principle of explicability: "operate transparently" must deal with what is meant under a transparent operation, what level of transparency and traceability we require. Due to the importance of these aspects, we have addressed these issues as part of our suggestions relative to Chapters II and III (in Section 3).Summarizing the above, we are of the view that the focus of this Draft and the future discussions about normative AI regulation should focus on the do no harm, transparency, and human agency principles and stick to realistic expectation.Critical concerns:Joining the underlying parts of the Draft, we would like to further highlight the importance of data privacy related considerations. This is one of the most important aspects both from a technical and from a sociological point of view.While we share HLEG's opinion on the importance of legal grounds in terms of AI data processing, the purpose limitation principle of the GDPR [Art.5 (b)] could have the same weight as well. Practically, this would mean that personal data gathered by AI systems should be collected for specified,

those eventual areas that are not overseen by human. We understand that basically there are three types of learning mechanisms: supervised, unsupervised and reinforcement learning. According to our current knowledge of state of the art AI solutions, neither of these represent a required level of autonomy or independence that could give rise to stand-alone liability of the AI, separated from its human supervisor.A further potential solution could be that certain AI operators/users could be obligated to conclude compulsory insurance packages, which, apart from the monetary advantages, could assist the greater public in acceptance of these systems.What could also strengthen the public trust in AI and the enforcement of certain regulations established in the later stages, is the setting up of an independent authority. Such authority could watch over the development and operation of AI systems. This is also something that could be considered when building up the main pillars of the AI regulatory environment.Data Governance and Respect for Privacy:We are of the view that not even the fairest core AI function will work properly if it is using and relying on incorrect, incoherent or inadequate data.Therefore, we believe that the applicability of GDPR should be emphasized in the Draft. GDPR applies to the processing of personal data in the context of the activities of a data controller or a data processor established in the EU. Accordingly, if the operator of the AI falls under the GDPR, it must operate the AI in accordance with the provisions of GDPR. Elaboration of AI specific cases and regulation may very well be necessary as an addition to GDPR itself.We completely agree that the requirements laid down by the GDPR must be guaranteed at all stages of the life cycle of data processed in AI system. However, we believe that the stage of data collection is one of the most important steps of the working mechanism of AI. Therefore, it is to be highlighted that the operators of AI solutions, as data controllers should be required to include protective measure to comply with the requirement of privacy by design.We believe that the core asset of AI is its prediction ability. Nonetheless, AI, as a predictive system has side effects – just like medicinal products may also have – when it comes to prediction. The data and the learned patterns / representations of AI models can jointly be capable of predicting properties of human subjects well beyond the scope of the original modelling, see for example the surveillance camera example mentioned above. Therefore, from a data protection point of view detailed notification has to be provided to the data subjects on the data processing including every possible outcome of the prediction mechanism, or if such advance warning proves to be intractable, at least at the point when the data and learned representations are being utilized for an additional or extended purpose. In this regard the national data protection authorities should be obliged to raise the data subjects' awareness not only in terms of the protection of their personal data, but in terms of predictions derived from the data they provided to the data controller in order to reduce the possible digital harm caused to the data subject.Non-Discrimination and Respect for (& Enhancement of) Human Autonomy:We

those eventual areas that are not overseen by human. We understand that basically there are three types of learning mechanisms: supervised, unsupervised and reinforcement learning. According to our current knowledge of state of the art AI solutions, neither of these represent a required level of autonomy or independence that could give rise to stand-alone liability of the AI, separated from its human supervisor.A further potential solution could be that certain AI operators/users could be obligated to conclude compulsory insurance packages, which, apart from the monetary advantages, could assist the greater public in acceptance of these systems.What could also strengthen the public trust in AI and the enforcement of certain regulations established in the later stages, is the setting up of an independent authority. Such authority could watch over the development and operation of AI systems. This is also something that could be considered when building up the main pillars of the AI regulatory environment.Data Governance and Respect for Privacy:We are of the view that not even the fairest core AI function will work properly if it is using and relying on incorrect, incoherent or inadequate data.Therefore, we believe that the applicability of GDPR should be emphasized in the Draft. GDPR applies to the processing of personal data in the context of the activities of a data controller or a data processor established in the EU. Accordingly, if the operator of the AI falls under the GDPR, it must operate the AI in accordance with the provisions of GDPR. Elaboration of AI specific cases and regulation may very well be necessary as an addition to GDPR itself.We completely agree that the requirements laid down by the GDPR must be guaranteed at all stages of the life cycle of data processed in AI system. However, we believe that the stage of data collection is one of the most important steps of the working mechanism of AI. Therefore, it is to be highlighted that the operators of AI solutions, as data controllers should be required to include protective measure to comply with the requirement of privacy by design.We believe that the core asset of AI is its prediction ability. Nonetheless, AI, as a predictive system has side effects – just like medicinal products may also have – when it comes to prediction. The data and the learned patterns / representations of AI models can jointly be capable of predicting properties of human subjects well beyond the scope of the original modelling, see for example the surveillance camera example mentioned above. Therefore, from a data protection point of view detailed notification has to be provided to the data subjects on the data processing including every possible outcome of the prediction mechanism, or if such advance warning proves to be intractable, at least at the point when the data and learned representations are being utilized for an additional or extended purpose. In this regard the national data protection authorities should be obliged to raise the data subjects' awareness not only in terms of the protection of their personal data, but in terms of predictions derived from the data they provided to the data controller in order to reduce the possible digital harm caused to the data subject.Non-Discrimination and Respect for (& Enhancement of) Human Autonomy:We

and how European companies, organisations, researchers, public services, institutions, individuals or other entities can successfully apply AI, would result in global competitive disadvantage and would naturally trigger the scaling down of European AI at a global level. Even if it remains to be seen how US and China for instance would regulate the functioning of AI, it is clearly visible that EU, due to historical and cultural reasons is more conservative and concerned and less risk-taking with AI. Europe's competitive disadvantage relative to the deployment of AI backed solutions may result in declining prosperity of European citizens and societies, and at the end, would trigger less bargaining power for EU on setting the global AI agenda.Therefore, it is essential that EU not just creates appropriate ethical framework, but one that enables that European AI will be competitive at global scale. We must therefore understand that our regulatory approach should— stick to minimum level of regulation;— enough risk taking in return of development;— leave behind the overprotecting approach;— create ways of "fast-track" development for AI stakeholders in case of voluntary and expressed partaking of individuals (just like trial application of drugs);— allow widespread usage of AI solutions protecting citizen's rights not by restricting the use cases of AI, but by strictly enforcing principles and technological solutions for privacy by design.Moreover, we believe that the Draft should underscore more the innovation aspect and elaborate further details on the risk-reward concept of the development and deployment of AI solutions, both at the subject and the developer level.We have to identify those regulatory issues in AI that we have not encountered so far, and we only have to deal with those aspects, instead of taking AI as a standalone "animal" and start drafting separate pieces of legislations onto its each and every aspect. Furthermore, AI as a software remains heavily exposed to its human creator and errors in the AI software made by humans should undergo the same software liability regime as other software. The defects of the hardware of the AI would also remain under general hardware malfunctioning rules.Accepting this approach would mean that a modern AI regulation should have strong data regulatory aspect, would be built on already existing liability regimes, and it would have a separate liability regime on the core essence of AI functioning. It is beyond doubt that at the end AI is something completely new in its core functioning, and that needs to be appropriately addressed.Enforcement of EU AI rules:In the age of globalization it is extremely hard to regulate and enforce EU regulations against web-based services and internet embedded interactions and dealings. Just consider the applicability of GDPR to companies (as data controllers or processors) seated outside of EU, yet subject to point 2 of Article 3 of GDPR.To date, we are not aware of any cases, moreover any technics how GDPR could be enforced by EU data protection authorities outside the soil of EU in lack of cooperation by the relevant country and/or the affected services provider.This fact may warn us that even if we overregulate the European way of AI, our citizens will be still exposed to non-regulated AI activities, and what is more important, that EU AI start-ups and companies will have

				explicit and legitimate purposes and not further processed in a manner that is not transparent and is incompatible with those purposes or represents a radical broadening of goals and scope (e.g., data collected for the legal purpose of ensuring security of a publicly accessible space should not serve as a ground for predicting personal traits and characteristics of a subject, etc.). With strong focus on purpose limitation, we could avoid inappropriate and overreaching usage of our personal data.Finally, let us also stress the growing concerns around the topic of consents by the data subjects. While fortunately the Draft put strong emphasis on the importance of informed consents in terms of AI usage, today's trends show that the real value of consents seem to be devaluing, as the general data protection awareness of users is yet to increase and consents are usually given without actual knowledge of the details of the data processing. Therefore, we suggest initiating discussions about eventual solutions that might be able to counteract the data subject's exposure and facilitate the appropriate information of data subjects about data processing, as well as to come up with legal intervention measures to prevent the monopolistic accumulation of data capital about EU subjects without any counter-balance of value caused by consents acquired practically "for free".	believe that it is important to underscore that the core functioning of AI works in a way that a predictive model, based on historical data makes/suggests decisions or predicts outcomes. In such a way, AI cannot be considered discriminative per se in connection with its inputs. If the output of an AI model is discriminative, that could mean two things: either data populated in the AI model was discriminative (biased by observing historical bias by negligence or design), or the software developer wilfully or negligently designed the objectives and boundary criteria for modelling in a biased or unfair manner. Thus, we propose to decompose "AI bias" to "data bias" and "modelling bias" respectively.	believe that it is important to underscore that the core functioning of AI works in a way that a predictive model, based on historical data makes/suggests decisions or predicts outcomes. In such a way, AI cannot be considered discriminative per se in connection with its inputs. If the output of an AI model is discriminative, that could mean two things: either data populated in the AI model was discriminative (biased by observing historical bias by negligence or design), or the software developer wilfully or negligently designed the objectives and boundary criteria for modelling in a biased or unfair manner. Thus, we propose to decompose "AI bias" to "data bias" and "modelling bias" respectively.	unbearable competitive disadvantage and will not be successful on the long run.So, we should not rule out to consider the complete ban on AI solutions that are not compliant with EU rules targeting EU citizens outside from the EU. Otherwise, the AI rally would be an uphill battle for EU companies and developers, as well as to consider legal means to appropriately rebalance the massive outflow of value, represented by our data capital to non-EU based providers and entities (the feasibility of collecting of "data customs payments" and enforcing technical solutions for privacy by design has to be thoroughly investigated).
Aleksander	Tarkowski	Fundacja Centrum Cyfrowe	We welcome the introduction of this important framework, which defines needed ethical standards for the development of AI technologies in Europe. Our comments mainly concern how these standards will interact with regulatory measures, and how to ensure they remain relevant and up-to-date despite rapid technological changes. On one hand, we see ethical standards and other forms of soft-law as having the advantage of being relatively agile forms for shaping technological development. On the other, we believe that regulation needs to be at least taken into account, and possibly implemented, if that is necessary to ensure that in particular fundamental human rights are respected. We understand the purpose and nature of the guidelines, and that in particular they are not meant to substitute regulation of AI. Still, it is difficult to fully evaluate these guidelines without being able to refer to regulatory recommendations. Especially that the document defines a relationship between ethical standards and regulation: "Trust in AI includes: trust in the technology, through the way it is built and used by humans beings; trust in the rules, laws and norms that govern AI [..]" (p. 2). Based on such a definition of the central concept of trust, we cannot answer the question whether AI will be trustworthy, without defining the legal / regulatory framework. p.2 - "it should be noted that no legal vacuum currently exists, as Europe already has regulation in place that applies to AI" - this crucial statement, which could determine HLEG's work on regulatory measures, is provided without any evidences. Regulation of AI, as well as definition of ethical standards should be	In the section on "Fundamental Rights of Human Beings", we believe that the freedom of speech, expression and information should be explicitly listed in section 3.2. We note the importance of highlighting these basic freedoms with regard to the current debate on the introduction of algorithmic filtering through the new proposed Directive on Copyright in the Digital Single Market. P. 11 - "Critical concerns raised by AI". Ethical guidelines should be based on a thorough risk assessment of AI technologies and their impact. We believe that the approach adopted in the guidelines is insufficient, as it presents a non-exhaustive list, on which the HLEG's members have not agreed upon. We welcome it as an initial analysis of the issue. A more robust mechanism is needed to identify all risks and concerns. We would like to point out in particular to the need of involving users of AI technologies in the risk assessment process. Participatory methods exist to support such processes, such as for example the citizen's panel methodology. THe work of the HLEG and the AI Alliance should be supported by the introduction of stronger participatory mechanisms that would engaged more broader groups of stakeholders and citizens.	We welcome an approach that achieves trustworthy AI through technical means, by applying the "Ethics & Rule of law by design" principle. The framework proposed in the document should be developed into a label or mark of quality. The recently introduced "Trustable Technology Mark" (https://trustabletech.org/) is an important project of this type, which could be used as a point of reference.	We welcome the introduction of the assessment list and continuing work on its operationalisation. Nevertheless, such a list will not become a useful tool for ensuring trustworthy and ethical AI without an established validation procedure, which in particular should depend on external review.	We would like to once again state the importance of developing a regulatory framework alongside these ethical principles. It is hard to establish the impact of these principles without reference to a regulatory framework, especially that the framework itself relates to those. We believe on one hand that ethical standards and other forms of soft-law have the advantage of being relatively agile forms for shaping technological development. On the other, we believe that regulation needs to be at least taken into account, and possibly implemented, if that is necessary to ensure that in particular fundamental human rights are respected. Having said that, the guidelines are an important step in developing a soft-law approach that is important with regard to emergent technologies. For this reason, this framework should not be "voluntary", and should guide all development of AI in Europe. Furthermore, the guidelines should be strengthened by introducing: 1) participatory mechanisms for the monitoring, review and updating of the guidelines; 2) a certification or a label for products that comply with the guidelines. With regard to the document "A Definition of AI: Main Capabilities and Scientific Disciplines", the purpose of the definition is unclear, despite the expansion of the definition. It is not clear to us why the definition expands on AI only as a scientific discipline, while omitting other important ways, in which the definition could be expanded. While there is a need to explain the understanding of AI as a technology and scientific discipline, it should not constitute the core of a definition that will be mainly

			based on a thorough review of regulation, and in particular on identifying existing legal gaps. In our opinion, the guidelines underestimate necessary adjustments to EU regulations that deployment of AI creates. There is a need for in depth discussion on the regulation of AI development. P. 6 - the authors argue that an ethical approach will allow a "unique brand of AI" to be developed in Europe. We welcome this ethics-based approach to AI design, but note at the same time the importance of ensuring an innovative and competitive business sector based on AI technologies. The guidelines should analyse, at least in broad terms, how these goals can be achieved in parallel. The document is described as a "living document", a "new and open-ended process of discussion". We welcome this ambition, and agree that the document should have a "living character". At the same time, we believe that the standards require a stronger participatory and governance mechanism to ensure representation of all voices and points of view. In particular, a monitoring mechanism for the implementation of the guidelines should be established.			used for policy making, strategic development and regulation. In our opinion, the term AI is very broad and without using more precise terms, makes it difficult to develop an approach to AI ethics and regulation. It is an almost catch-all term that covers both existing implementations of AI and "advanced robots". One specific, and precise term that we would like to see included in the definition is "algorithmic decision-making". The use of this term allows debates on ethics and regulation to focus on the issue of how a subset of AI technologies and their implementations makes decisions about citizens and all aspects of social, political and economic life in Europe. For the application of this concept to an analysis of deployment of AI technologies, see the recent report "Automating Society – Taking Stock of ADM in the EU", prepared by AlgorithmWatch (http://www.algorithmwatch.org/en/automating-society/). Implementations of such algorithmic systems create a specific set of challenges, risks and potentials that should be addressed. We suggest adding it to the part "Other important AI notions and issues".
Valentine	Goddard	AI Impact Alliance (AIIA)	The challenge lying ahead is mainly in seizing a growth opportunity without being derailed by its complexity, the speed at which it is going, or the pressures of the market. When facing these challenges, let's keep in mind that Civil Society Organizations (CSOs) have the flexibility and network required to navigate disruptive societal changes and are active and precious partners of public institutions, particularly in this period of history. Not only do CSOs have the critial role of maintaining political and social balance, they are the trusted allies of citizens. They are educators, mediators and social dialogue experts. Further exploring innovative methods of civic engagement and deliberation is pivotal in improving a lack of trust in public institutions, and technology.The current draft refers to "all relevant stakeholders (…) encompassing companies, organisations, researchers, public services, institutions, individuals or other entities". the AI Impact Alliance (AIIA) recommends that the European Commission officially recognize Civil Society Organizations as an important partner and, as such, be specifically listed and named as stakeholders in the Ethical Guidelines. Secondly, AIIA agrees that "AI is key for addressing many of the grand challenges facing the world, such as global health and wellbeing, climate change, reliable legal and democratic systems and others expressed in the United Nations Sustainable Development Goals." For that reason, we recommend the use of strongly suggestive terminology, throughout the Ethical Guidelines, to ensure that the ethical framework not only protects existing human rights, but also helps steer the use of AI to achieve the Sustainable Development Goals. Although keeping this goal in the introduction gives a "guiding" value to the achievement of Sustainable Development Goals, including it in the body of the Ethical Guidelines would give it more weight. The same suggestion goes for this part of the text: "its development,	2.0 Non-Technical MethodsAs we move forward, we need to remember that our laws are written representations of societal values. Civil Society Organizations are – among other important roles – significant and pivotal communication channels between citizens and public institutions. Therefore, we believe they are best placed to facilitate an inclusive dialogue on AI's impact on society, bring forth the results of their findings and have an impact on regulatory innovation regarding the deployment of AI in society.Establishing international principles and guidelines on AI will require considerable collaboration between different countries, as well as across sectors, which do not always share the same goals nor motivations. Efforts to reach a consensus are further hindered by fierce economic competition, which can undermine the effectiveness of initiatives to adapt our normative structures to a rapidly transitioning society. In fact, in order for standards to be effective and implementable, we argue that the consensus that global partners need to reach will be attained faster by empowering Civil Society Organizations to lead these efforts. Recent research (Jordan, C. (2018) International Policy Standards: An Argument for Discernment) has shown that in some high profile efforts in the financial sector, because of the stakes at play, efforts in setting and enforcing international standards that are led by the private sector risk collapsing. This reinforces the argument that Civil Society Organizations have a vital role to play in advancing towards a consensus on pressing issues related to AI Governance.Unfortunately, despite a few exceptions, Civil Society Organizations have yet to receive the support they need to effectively be part of the discussion on AI governance. Public-Private partnerships have already initiated national and international workgroups towards achieving consensus relative to AI governance. However, in the hallways, there are shared concerns about	In the lexical guide, AIIA recommends defining Digital Literacy as it is often narrowed to teaching citizens how to code. However, the decisions to be made by the citizenry in the age of AI are much more important. In fact, there is a broad need across sectors and disciplines to understand the ethical and societal implications of AI, as well as the political choices that should be decided by voters in a democratic system. Bridging the digital gap also involves that universities from around the world, including the Global South, can participate in AI conferences and workshops, rub shoulders with other scientists, and share the complimentary expertise for eradicating poverty, for example. Their conversations, sharing of ideas, research results and network will elevate our collective intelligence and contribute to making AI a tool that enhances humanity's well-being.

deployment and use should respect fundamental rights and applicable regulation, as well as core principles and values, ensuring an "ethical purpose" ". We recommend adding to this: " and help steer the use of AI towards the achievement of Sustainable Development Goals". In short, a proactive promotion of human rights and Sustainable Development Goals will maximise the benefits AI can bring to the society and the world we live in.

the ethical implications of the private sector leading the development of ethical and normative frameworks on AI. Public institutions have the responsibility to ensure that ethical and normative frameworks development is not lead by the private sector, or those who have more technical expertise. It wouldn't be forward thinking to argue that technical expertise justifies the absence of organizations that aren't as digitally savvy. It would also contradict the predominantly accepted necessity for multidisciplinary perspectives in order to grasp the many nuances of AI implementation into society (ex: accountability, acceptability).In short, Civil Society Organizations must have all the necessary resources to understand the issues and be empowered to participate in the process of developing ethical and normative frameworks. The Ethical Guidelines have the opportunity to make it a priority to address the current imbalance.More specifically in the following sections:Regulation:AIIA would like to underline the fundamental importance of the participation of Civil Society Organizations in the on-going process of determining the need to "revise, adapt or introduce such regulations" pertaining to AI's ethical and societal implications, from its initial development stages to its deployment into society. The capacity for Civil Society Organizations to participate in this process will be made possible with funding and education programs geared towards the empowerment of organizations that work with citizens, that the HLEG has justly addressed further in Education and awareness to foster an ethical mind-set. Standardization:The existing workgroups on ethical standards related to AI are not made accessible to Civil Society Organizations, very often by means of cost of membership, or travel cost to participate in such meetings. "Using agreed standards for design, manufacturing and business practices" may well not achieve its goal to "function as a quality management system for AI offering consumers, actors, and governments the ability to recognize and reward ethical conduct" if for example Civil Society Organizations of high level expertise on consumers rights are not able to participate in the development of such standards. The Roundtable for Sustainable Biomaterials (RSB) and the International Development and Research Center (IDRC) for example fund the participation of stakeholders from the Global South to ensure that its multistakeholder and global aspirations are credible. Codes of Conduct:The following recommendations seem to be in line with the overall objectives of the Draft of the AI Ethics Guidelines. 1- Adding incentives to collaborate across sectors using tax breaks, much like a research and development tax, and the taxation of profitable uses of data to fund public services and AI research and development that serve the public good.2- Participants at the AI on a Social Mission (AIOASM) conference also supported the implementation of a Social Impact Index (SII). A SII would assess the value a company contributes to society, as a guide for investors and public funders. 3- Similarly, support organizations who develop and implement Social Return On Investments (SROIs) criteria. Incorporating them into

their return on investment reports should be incentivized. Governments could evaluate, through the use of algorithms, whether a company rates highly on an SII, and make it conditional to allocating public funding.These as well as other recommendations were made by the participants at AIOSM in 2018, the first conference on AI to have more than a quarter of its participants come from the non-profit sector. Civil Society Organizations (CSOs) were inspired by the positive potential of AI for their missions, and learned about risks. It was also a groundbreaking opportunity to represent CSOs voices in policy recommendations. Other participants included AI and data scientists, policy makers, members of the government, social innovators, social entrepreneurs, jurists, ethicists, students and researchers in various fields, and lay citizens interested in AI.More recommendations can be found in the Publications section of AIIA's website (www.allianceimpact.org). Education and awareness to foster an ethical mind-setKeeping citizens informed can be done in many different ways. The chosen methods and the involved partners will have an impact on the level of citizen trust in AI and their governments. Working with Civil Society Organizations who have already established a trust relationship with the citizens they serve is a winning collaboration worth underlying in the Guidelines.Indeed, Civil Society Organizations are the best digital literacy agents if such is the vocation of their organization. They are multipliers and information disseminators on the ethical and social impact of AI. As citizens gain understanding of the risks and benefits of AI, they will become less resistant to change and more enthusiastic to embrace change with bottom-up designed solutions and inclusive growth. AIIA believes it is crucial to avoid causing fear or intimidating citizens and recommends supporting efforts to foster interest and understanding of AI's potential risks and benefits in all citizens. As citizens discover the potential benefits of AI, they will develop the desire to see AI used for beneficial purposes, and with the support of CSOs and other stakeholders, engaged citizens can become part of those helping steer the development of AI towards the achievement of SDGs. Art may be very useful in these times of change, as a tool for education, and for social and cultural mediation. In fact, outreach and educational empowerment can be done through independently funded public art, and other multisectoral and multidisciplinary social dialogue initiatives (as fittingly recommended by the HLEG in "Stakeholder and social dialogue").Diversity and inclusive design teamsThe social deployment of AI is considered by many as posing the risks of increasing inequalities. In that context, AIIA recommends specifically listing socio-economic status in this section. AIIA also recommends promoting not only the diversity in "the teams that design, develop, test and maintain these systems reflect the diversity of users and of society in general." but also in the teams that contribute and oversee the governance of AI through regulation, standards, business and governance models and public policies.

Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Anonymous	Anonymous	Anonymous	Une IA digne de confiance ne peut se faire que dans le respect des règles, lois, par l'évaluation positive et corrective des normes à chaque fois qu'un correctif est signalé. l'évolution dans les respect le plus juste des droits souverains réservés à chaque personne ne pourra se faire que dans une bonne lisibilité des perspectives et des développements.EX- Fonctionnaire Territorial Titulaire Assermentée -expérimentée au développement de l'IA par la gestion des biais de contraintes totales (humaine et technique)Démissionnaire! à la suite d'un accident reconnu imputable au service après:-Piratage intentionnel depuis mon poste de travail de type ATP/ TOR, (preuves tardives)-Détournement de documents, base de données de 10 ans ect., perte de prérogatives.-Echanges de données de santé entre assureurs privé/professionnel (sous forme de partenariat, pas informée!)-Violation de consentement et d'identité,-Utilisation des données à des finalités autres qu'en SST / instruire le changement par test de la mobilité forcée.- Prévention primaire discriminante en santé au travail, - Lean mangement+ ingénierie sociale en réseau = sept années de vie extrêmement difficiles, sans droits fondamentaux préservés.-Sans statut à ce jour suite à cet enchaînement en liens de responsabilité élargie à une chaîne d'acteurs responsables dans un développement rendu illisible par cloisonnements étanches au droit à l'information pour une l'équité et pour la préservation des droits fondamentaux: prévention, santé, travail, vie privée, retraite, - étude d'impacts RGPG en attente. - Impuissance des autorités et de la justice à la résolution de ce type situation.- Avis sur les retours des régulateurs Français: CNIL Pack assurance 2014- RGPD et ACPR (fraude et blanchiment)- Prospectives sur l'ORSA -NPA5- DDA- EIOPA - OIT-OCDE- - Convention de Budapest pour le respect des droits fondamentaux des produits à doubles sens créés.-Droit à l'informationGestion du consentementMaîtrise des Données de santé et sensibles- établissement de la limite privé/professionnel dans la gestion des données entre assureurs.Respect de la vie PrivéeDroit au travail, social, bien être, justice, Responsabilité des préjudices, réparations des risques provoqués Éclaircissements des méthodes de gestion utilisées par les actuaires, assureurs, courtiers, mandataires, intermédiaires au regard des assurés sur le risques non maîtrises par la GED, les AAU, la dématérialisation, la conception des interfaces de logiciels, les failles technique et humaines, les exigences d'implémentations relatives aux normes, certifications, agréments, règlements, lois, droits.	L'être humain doit rester le centre des fondements de la prévention en SST pour lui et pour les autres- L' exclusion de ce principe de base est détourné par le jeu de la prévention" Primaire", à visée discriminatoire. Les plans, économiques, sociaux, santé et éthiques s'en retrouve être directement impactés à plusieurs niveaux. Ces principes ne sont pas conformes aux fondamentaux lorsque les droits sont détournés ils rendent ce type de gestion de la santé au travail très dangereux pour les personnes profilées.Les épuisements professionnels, Burn-Out- Suicides sont liés à ce type de concepts déshumanisés, la chaîne d'acteurs des traitements utilisent les manquements aux obligations de sécurité et de prudence inscrits dans la loi pour utiliser les données à d'autres finalités ceci vient gonfler les chiffres de l'absentéisme en collectivités publiques. La base de ce type de traitement ne répond pas au code du travail ni aux normes de sécurité (L 4121.1.2) exigéLes droits fondamentaux se trouvent directement impactés.	La charte: dignité, libertés, égalité et solidarité, droits des citoyens et justice au traité de l'EU doit être mieux respectée dans ce type de traitements. La discrimination induite est à la limite du supportable.Les droits: civil, politique, économique et social se retrouvent être très limités.Ces traitements sont indignes de confiance.Le secret médical est bafoué, les données de santé ne peuvent par respecter les cadres légaux prévus, les excès sont inadmissibles en gestion de données sensibles.L'incertitude réglementaire ne permet aucune lisibilité: une plainte classée sans suite par manque de preuves au TG I + 1 référé suspensif + 4 requêtes auprès du juge de l'excès de pouvoir TA= frais de justice sans droits reconnus, il n'y a euaucune explications à la suite de cette discrimination, c'est inacceptable.Développer de l'IA sur la base de ces concepts aussi peu aboutis mettent les personnes en grave danger.	Les valeurs fondamentales ne peuvent être garanties qu'avec le retrait de certains actes administratifs unilatéraux (AAU), par ce retrait, une décision créatrice de droits à la demande d'un tiers si elle a été prise illégalement ou lorsque la liberté est altérée par des décisions prises sans consentement préalable et ayant provoqué des préjudices graves devrait rester lisible à l'information et accessible à tous . Responsabilités:Dans la disruption et la dématérialisation des dossiers et lors des créations d'interfaces propres à certains logiciels, des biais fondamentaux relatifs à la santé, la prévention, au travail, la retraite et la vie privée, ont été créés ces traitement devraient pouvoir être remis en cause au titre de la transparence et la traçabilité. Si les actes administratifs unilatéraux par leurs résultats provoque une QPC (Conseil Constitutionnel), elles devrait être traitées plus équitablement et rester accessible à tous. La Charte EU, concernant les personnes faisant l'objet de discriminations par le traitement de leurs données à caractère sensibles en santé, validées par des conventions AAU puis couvertes par décret d'application peuvent se révéler être des atteintes graves aux droits fondamentaux, en amont des collectes: lors des conventions dématérialisées à l'application du décret et après, à l'heure des évaluations, il conviendra de vérifier si elles respectent le droit à l'information pour une bonne maîtrise et la gouvernance des données personnelles, cette logique va dans le sens de l'étique et le respect des droits.http://www.vie-publique.fr/decouverte-institutions/institutions/approfondissements/abrogation-retrait-actes-administratifs-unilateraux.htmlCertains logiciels de gestion des AT et Maladies étaient non sécurisés, ils ont été accessibles pendant plusieurs années à des personnes non habilités (non administratrices de ce droit) , les mots de passe ont été déclarés vulnérables, ces failles techniques ont permis des saisies en sous estimant les accidents imputables aux services par soustractions ou modifications des fiches personnelles pour ne pas être comptabilisés en imputabilité, les responsabilités de ces accidents sont ainsi rendues illisibles, les dossiers ne sont pas traités à temps, les personnes impactées se retrouvent dans des conditions de vie indignes, les responsabilités sont diluées, les droits fondamentaux se retrouvent tronqués. Le rapport de l'IGAS mentionne des incohérences non résolues.Il est a noter que certains correctifs ont été apportés à compter du mois de juin 2018. La détection de l'usurpation des données personnelles est difficile a appréhender, il s'en suit directement la violation de consentement, puis une utilisation à d'autres finalités. La violation du consentement chez une victime entretient le camouflage des mensonges d'un fraudeur, cela, produit des faux en écritures publiques y compris devant juridictions, elles renforcent les contournements à la sécurité, isole la victime en violant les codes éthiques de différentes professions ordinales (y compris celui des médecins).La vigilance s'impose en	Les biais qui m'ont été imposés depuis fin 2011 sont inacceptables et humainement dégradants, ils provoquent l'exclusion sociale et des atteintes graves, la perte de confiance dans les professions ordinales, la justice (médecins, avocats..).Les actions en cybercriminalités restent souvent impunies,le RGPD n'est pas respecté: fin de non recevoir!Nombreux sont les documents inaccessibles par manque de droit d'accès, La discrimination n'est jamais expliquée, elle sert juste de contre mesure à une prévention servant à écarter les profilés du vrai concept Prévention.Les excuses ne sont pas présentées, les ruptures sont totales et incomprises.La perte de chance n'est pas reconnueLes actes producteurs de droits négatifs demanderaient à être retirés en urgence.Ces traitements laissent des impacts néfastes aux humainsMerci d'apporter rapidement les corrections lisibles à ces failles.(A disposition pour détails complémentaires)

particularité dans les schémas: de prévention et d'évaluation, car si cette dernière est erronée, mal mesurée, non corrigée, c'est bien à ce moment précis que le déplacement des responsabilités dans le traitement est altéré. Le biais est crée, le principe éthique d'un médecin peut basculer sans qu'il ne s'en rende compte et celui de la victime également par l'entretien de la confusion sur une chaîne d'acteurs, la perte de confiance devient totale, il ne reste alors: que la solution de replis, pour essayer de comprendre les carences à corrigerSi une mauvaise évaluation reste définie comme prioritaire lors d'une prévention, elle produira un résultat binaire qui deviendra un bien à double sens qui se révélera être contraire aux objectifs initiaux.gestion du consentement:Lorsque la préservation de l'autonomie d'une personne est subordonnée par des expertises sous contraintes, exigées à la suite des violations graves, elles sont réalisées par des experts manquant totalement d'impartialité, payées par l'assureur lui même assurant le suivi du profilé depuis l'origine de la violation pour évaluer les risques provoqués par le concept: les contradictoires sont alors impossibles. l'IA doublée de l'action humain d'un autre age est non éthique, les bases juridiques sont contournées dès le départ du schéma.Gouvernance des données: pas de lisibilité - accès très difficile- pièces produites non accessibles, pourquoi?, Les principes de justice deviennent inapplicables car les effets induits par l'IA sont inexpliqués, difficilement explicables ou encore incompris, même avec des normes élevées, les certifications, les agréments permettent difficilement l'exercice des responsabilités humaines, elles sont noyées, contournées voir gommés dans une chaîne d'acteurs cloisonnés, les conséquences aux fraudes deviennent encore plus nuisibles aux humains face aux machines froides qui en prennent le relais.La gestion d'un profilage et d'un consentement ne doivent-ils rester incompatibles en santé après ce type de biais provoqué ?.(il faudra reposer la question aux concepteurs pour mieux étayer la sécurité des personnes)La détection de la fraude, du blanchiment de capitaux ou du financement du terrorisme, n'ont lieu que bien plus tard, des années après, pendant tout ce temps, la société vous a exclue, ignorée, bannie et vous subissez toujours des contres mesures discriminantes: droits non reconnus, frais de justices, fins de de non recevoir, intimidations, dégradations, intrusions en propriété privée. Toutes ces actions servent à camoufler les contournements d'éthiques et les nombreuses violations de droits, de non conformité avant les détections de fraudes, pendant ce temps vous n'avez aucune protection juridique, vous êtes sans statut, vous ne rentrez plus dans un aucun concept, vous êtes exclu de droits fondamentaux.

Pam	DIXON	World Privacy Forum	World Privacy Forum 3 Monroe Parkway Suite P #148Lake Oswego, OR 97305USA 30 January 2019Re: Comments of the World Privacy Forum regarding the European Commission's High Level Expert Group on Artificial Intelligence Draft Ethics Guidelines for Trustworthy AIThank you for the opportunity to provide comments regarding the Commission's High Level Expert Group on Artificial Intelligence Draft Ethics Guidelines for Trustworthy AI, Working document for stakeholders' consultation, available at https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=57112. The context of our comments is that of privacy, and specifically, privacy seen in its relationship as an important aspect of human autonomy and other human rights. The World Privacy Forum is a non-profit public interest research group that focuses on consumer data privacy issues, including those relating to emerging technologies, identity, data brokers, AI, health, and other topics. WPF is a non-political, non-partisan organization. WPF works exclusively on privacy and data protection, and is one of the only US NGOs that focuses on objective research so as to produce fact-based consumer data privacy work. Our research, testimony, consumer education, and other materials are available on our webpage, www.worldprivacyforum.org. Regarding AI and ML, WPF researched and wrote a major report about predictive analytics and privacy, which is contextualized in the US legal framework. Additionally, we have spent substantive time researching and writing about biometrics, which is an important subset of AI and ML. Our research on India's Aadhaar biometric ecosystem, impacting over a billion people, was cited twice in The Supreme Court of India's landmark privacy decision. The theme of our biometric work in India, and one that the Supreme Court addressed, was our call, based on the facts from our findings, is that biometrics must do no harm, and must create a public good. I. Comments on the PrinciplesWe find ourselves aligned with almost all of the Expert Group's principles. We offer no comment on weapon systems, as this is not in our mission and purpose. A. Support for Do No Harm / Create a Public Good We do, though, strongly support all of the other principles. We want to particularly endorse the importance of the Do No Harm concept, as well as the Do Good concept. After much thought regarding the positive statement of providing benefit with AI, we have come to appreciate the phrase "create a public good." We submit this for your consideration. We believe that the concept of Do No Harm and its corollary, Create a Public Good, is the correct bedrock for AI and ML principles. B. Comments on Section 3.4 We support section 3.4. However, we request that gender is specifically included. We have noticed, particularly in the Global South, that gender inequality and discrimination is a meaningful problem, and deserves to be brought forward in particular so there is no mistaking its importance. In some jurisdictions, women constitute a vulnerable population. II. Comments on Biometrics, Identity, and AI/ML Identity is a data-rich key that acts to unlock all levels of the emerging digital ecosystem. All forms of ID carry some risk, but digital forms of ID, or "dematerialized ID," cuts across all sectors

and generates particularly copious data about people, their behaviors, financial status, associates, and potentially even political and religious views. Over time, distinct patterns emerge from the data and have in the past created new kinds of risks for individuals and groups. As the world is becoming increasingly digitized, we can expect challenges in the identity space to grow apace unless proactive attention is given to identifying and mitigating the risks. The principles mention identity and biometrics in section 5.1, Identification without Consent. This discussion is correct, but incomplete. It does not capture the full scope of the issue. We draw your attention to two key case studies in biometrics, those in which government is a key actor. India, which has provided the world's most significant case study on the implementation of nation-wide biometric systems in voluntary and non-voluntary environments, provides important lessons. As mentioned earlier, WPF researched the Aadhaar ecosystem extensively in the field, and wrote a large research report on the system. Our research and policy analysis was cited twice in the Supreme Court of India's landmark Aadhaar case, in 2018. India went from adding its first voluntary enrollee in its Aadhaar biometric ID program in 2010, to boasting more than 1 billion enrollees in 2016. In order to allow for innovation, growth, and modernization, privacy and data protection regulations were eschewed in favor of technological advancement and modernization of the governmental, financial, health and other sectors. The Aadhaar digital identity ecosystem was intended to act as an identity key for the poor and to allow for unfettered, frictionless delivery of subsidies. The vision was well-meaning, but the system suffered from multiple challenges, including security breaches, that caused the entire system to be brought into question. Ultimately, the system was sharply curtailed by the 2018 Aadhaar Supreme Court of India decision. One notable challenge the system experienced was significant mission creep, which caused a lack of user trust in the system over time. Instead of just being used for delivery of subsidies, it became increasingly difficult to get paid, receive pensions, file taxes, bank, or get health services in India without an Aadhaar ID. As the Aadhaar become used more widely, Aadhaar also went from being a voluntary system to a mandatory system. Three factors: the lack of stakeholder input, mission creep, and eventually a loss of user trust in the system, are what truly caused the curtailment of Aadhaar. The lack of policy and governance allowed these problems to persist without being addressed. Currently, Kenya's national identity system is showing early warning signs of a system exemplifying what we now know are very poor identity and data practices. Kenya's government has added amendments to existing identity legislation enabling the collection of DNA from its citizens and foreign residents. The DNA is planned to be put in a centralized national database, and used by the government for multiple purposes. No collection has occurred yet, but already, unrest and deep concern over the potential for serious abuse of a centralized DNA database has arisen. A key difficulty is that Kenya has passed legislation allowing

the DNA collection, but it has not yet passed overarching data protection legislation that would protect individuals from abuse of the identity data, or provide avenues for redress if harm has occurred. The stage is set for significant harm to develop in respect to Kenya's identity ecosystem. Unless the government of Kenya enacts significant baseline legislative and policy protections incorporating protections in place prior to the collection, creation, or use of a central DNA registry, then the system is likely to cause potentially profound harms. Aadhaar has already shown us where the end stages of centralized biometric identity database deployments are, what they look like, and how they operate. The lessons are already there, including the loss of trust the Aadhaar system experienced and the harm Aadhaar enrollees experienced. There is no reason to repeat these kinds of mistakes in Kenya, or elsewhere. Our hope is that the Guidelines will directly address the biometrics issue apart from just consent. Europe already has baseline data protection and privacy legislation in place, so some issues will be improved as a matter of course, particularly in the commercial sector. But there is a great deal of room for difficulty in government uses of biometrics, where it is much more difficult to see a pathway to meaningful consent. What are the guidelines that can address these issues? We believe the Expert Group can find a way to address this. Ideally, the Guidelines will have global impact. Given this, it is particularly important that the large mandatory biometrics use case problems are addressed, as non-EU countries need guidance regarding commercial as well as government uses. III. Comments on Tension points in AI and Machine LearningArtificial Intelligence and machine learning techniques have matured considerably in the past decade, affording new insights into data across multiple disciplines. Different flavors of AI exist: Convolution Neural Networks, Markov Models, Ensemble Methods, Deep Learning, Bayesian Belief Nets, Statistical Models. These models have different levels of explainability; there are some interpretable models, some models have the so-called "black box" which can be impenetrable. For some models, modified deep learning techniques can learn explainable features. It is crucial in policy discussions to distinguish between AI models and their differing levels of explainability. Much attention in the past few years has been given to a variety of tension points in AI, for example, the lack of transparency of the "black box." However, additional tension points exist, and should be treated just as thoughtfully. Fairness, transparency, accountability, and good governance around uses of AI and multiple other aspects of AI are among key aspects to include in any principles and policies regarding AI. The Guidance has done an admirable job of incorporating much nuance around these issues. We would like to pause here and in addition to supporting the Guidelines on these topics, also support Japan's AI Guidelines, which in 2018 are now a completed draft after substantive multistakeholder deliberation. The guidelines are thorough, fair, and balanced.We would like to discuss two tension points in particular. That is, input risks, and risks regarding interpretation and use of results. We focus on these two areas here. IV.

Comments Regarding inputs/data sets risksAI analysis is a data-intensive discipline, requiring abundant input factors ranging from raw data sets to algorithms, and in some cases, categorizations or scores based initially on raw data sets, a full accounting of the privacy risks associated with input factors is important. First, data sets must be available to use; second, data sets must be appropriately cleaned and prepared for use; and third, the data sets must be appropriately matched to the intended inferences or goals sought from the analysis. These are among the baseline considerations for data sets, understanding that many more considerations exist. Among these considerations includes potential issues relating to data sets that are derived directly from or about individuals or groups of individuals, or in some cases data sets that while not directly derived from or about individuals, can be used to create inferences about individuals or groups of individuals.Regarding algorithms or scores/categorizations used as input factors for AI analysis, a primary consideration (beyond ethical data use and the need for privacy assessments for enhanced risks) is that many of these types of input factors can be proprietary in nature. Given that some AI analysis utilizes numerous algorithms as input factors, proprietary algorithms could pose obstacles for AI use across industries or sectors over time, as well as pose substantial challenges to transparency, fairness, and interpretation. We mention data brokers here as an important category to think about. While data brokers are not as extensively operating in Europe, they are operating in other jurisdictions, and this has impacts on AI inputs and fairness Please see our report, The Scoring of America for many specific details of what this is, how it happens, the products/services in this space, and an analysis and recommendations for solving the problems. The issue of secondary use of data, and particularly secondary use in AI systems, is important to resolve. V. Comments Regarding interpretation and use of AI outputsHow to interpret the results of AI analysis also needs specific guidance. Interpretation should occur within an understandable, specific context and should be carefully constrained and defined. AI model results are only as predictive or as fair as the score model or models, the factors used in that model, and the training and fit of that model to the task or problem it was meant to solve for, among other factors. However, much interpretive nuance is easily lost when an AI model results in a simple numeric score. A simple score can be deceptively complex to interpret; models can be over or under fit, creating potentially significant discrepancies in results. Over-fitting arises when an algorithm is trained to perform very well on an existing set of data, but has been tailored so well to that data set that it can behave erratically or incorrectly outside of the specific scenario it has trained for. When a predictive model assigns a value or a range to a person, for example, a risk score, the model used to create that value must be transparent, accurate, reliable, and kept up to date. The numeric range for interpreting the result (such as a score) should be well-quantified, and the results validated.Without these protections, even the best and most predictive model can be interpreted improperly, to potentially

negative consequences.Currently, very little governance exists around the interpretation and use of specific AI results. It is an area particularly well-suited for further work. Governance models can be used to address the numerous contextual issues that arise in the area of use of AI scores or models. VI. Conclusion The successful design of principles for AI and ML must be effective today, and effective in the future. This is one of the great challenges in composing guidelines on AI and ML today; we are essentially peering into the future and working to anticipate risks in order to mitigate those risks. We offer one idea here: whenever possible, use unambiguous case studies to guide where the harms are. We are certain that India's Aadhaar system was deeply problematic. We are certain that as other national, large systems use or propose to use biometrics - including DNA — that they need baseline legal and policy protections in place, as well as ethical guidance. We are also certain that gender is an extremely important aspect of discrimination that occurs in AI and ML. There is robust technical support for this, including in the field of biometrics. And finally, after our fieldwork in India, we have come to deeply understand that AI and ML must do no harm and must create a public good. We again voice our strong support the inclusion of these principles and ideas in the draft. Respectfully submitted, S/Pam Dixon Executive DirectorWorld Privacy Forum World Privacy Forum 3 Monroe Parkway Suite P #148Lake Oswego, OR 97305USA 30 January 2019Re: Comments of the World Privacy Forum regarding the European Commission's High Level Expert Group on Artificial Intelligence Draft Ethics Guidelines for Trustworthy AIThank you for the opportunity to provide comments regarding the Commission's High Level Expert Group on Artificial Intelligence Draft Ethics Guidelines for Trustworthy AI, Working document for stakeholders' consultation, available at https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=57112. The context of our comments is that of privacy, and specifically, privacy seen in its relationship as an important aspect of human autonomy and other human rights. The World Privacy Forum is a non-profit public interest research group that focuses on consumer data privacy issues, including those relating to emerging technologies, identity, data brokers, AI, health, and other topics. WPF is a non-political, non-partisan organization. WPF works exclusively on privacy and data protection, and is one of the only US NGOs that focuses on objective research so as to produce fact-based consumer data privacy work. Our research, testimony, consumer education, and other materials are available on our webpage, www.worldprivacyforum.org. Regarding AI and ML, WPF researched and wrote a major report about predictive analytics and privacy, which is contextualized in the US legal framework. Additionally, we have spent substantive time researching and writing about biometrics, which is an important subset of AI and ML. Our research on India's Aadhaar biometric ecosystem, impacting over a billion people, was cited twice in The Supreme Court of India's landmark privacy decision. The theme of our biometric work in India, and one that the Supreme Court

addressed, was our call, based on the facts from our findings, is that biometrics must do no harm, and must create a public good. I. Comments on the PrinciplesWe find ourselves aligned with almost all of the Expert Group's principles. We offer no comment on weapon systems, as this is not in our mission and purpose. A. Support for Do No Harm / Create a Public Good We do, though, strongly support all of the other principles. We want to particularly endorse the importance of the Do No Harm concept, as well as the Do Good concept. After much thought regarding the positive statement of providing benefit with AI, we have come to appreciate the phrase "create a public good." We submit this for your consideration. We believe that the concept of Do No Harm and its corollary, Create a Public Good, is the correct bedrock for AI and ML principles. B. Comments on Section 3.4 We support section 3.4. However, we request that gender is specifically included. We have noticed, particularly in the Global South, that gender inequality and discrimination is a meaningful problem, and deserves to be brought forward in particular so there is no mistaking its importance. In some jurisdictions, women constitute a vulnerable population. II. Comments on Biometrics, Identity, and AI/ML Identity is a data-rich key that acts to unlock all levels of the emerging digital ecosystem. All forms of ID carry some risk, but digital forms of ID, or "dematerialized ID," cuts across all sectors and generates particularly copious data about people, their behaviors, financial status, associates, and potentially even political and religious views. Over time, distinct patterns emerge from the data and have in the past created new kinds of risks for individuals and groups. As the world is becoming increasingly digitized, we can expect challenges in the identity space to grow apace unless proactive attention is given to identifying and mitigating the risks. The principles mention identity and biometrics in section 5.1, Identification without Consent. This discussion is correct, but incomplete. It does not capture the full scope of the issue. We draw your attention to two key case studies in biometrics, those in which government is a key actor. India, which has provided the world's most significant case study on the implementation of nation-wide biometric systems in voluntary and non-voluntary environments, provides important lessons. As mentioned earlier, WPF researched the Aadhaar ecosystem extensively in the field, and wrote a large research report on the system. Our research and policy analysis was cited twice in the Supreme Court of India's landmark Aadhaar case, in 2018. India went from adding its first voluntary enrollee in its Aadhaar biometric ID program in 2010, to boasting more than 1 billion enrollees in 2016. In order to allow for innovation, growth, and modernization, privacy and data protection regulations were eschewed in favor of technological advancement and modernization of the governmental, financial, health and other sectors. The Aadhaar digital identity ecosystem was intended to act as an identity key for the poor and to allow for unfettered, frictionless delivery of subsidies. The vision was well-meaning, but the system suffered from multiple challenges, including security breaches, that caused the entire system to

be brought into question. Ultimately, the system was sharply curtailed by the 2018 Aadhaar Supreme Court of India decision. One notable challenge the system experienced was significant mission creep, which caused a lack of user trust in the system over time. Instead of just being used for delivery of subsidies, it became increasingly difficult to get paid, receive pensions, file taxes, bank, or get health services in India without an Aadhaar ID. As the Aadhaar become used more widely, Aadhaar also went from being a voluntary system to a mandatory system. Three factors: the lack of stakeholder input, mission creep, and eventually a loss of user trust in the system, are what truly caused the curtailment of Aadhaar. The lack of policy and governance allowed these problems to persist without being addressed. Currently, Kenya's national identity system is showing early warning signs of a system exemplifying what we now know are very poor identity and data practices. Kenya's government has added amendments to existing identity legislation enabling the collection of DNA from its citizens and foreign residents. The DNA is planned to be put in a centralized national database, and used by the government for multiple purposes. No collection has occurred yet, but already, unrest and deep concern over the potential for serious abuse of a centralized DNA database has arisen. A key difficulty is that Kenya has passed legislation allowing the DNA collection, but it has not yet passed overarching data protection legislation that would protect individuals from abuse of the identity data, or provide avenues for redress if harm has occurred. The stage is set for significant harm to develop in respect to Kenya's identity ecosystem. Unless the government of Kenya enacts significant baseline legislative and policy protections incorporating protections in place prior to the collection, creation, or use of a central DNA registry, then the system is likely to cause potentially profound harms. Aadhaar has already shown us where the end stages of centralized biometric identity database deployments are, what they look like, and how they operate. The lessons are already there, including the loss of trust the Aadhaar system experienced and the harm Aadhaar enrollees experienced. There is no reason to repeat these kinds of mistakes in Kenya, or elsewhere. Our hope is that the Guidelines will directly address the biometrics issue apart from just consent. Europe already has baseline data protection and privacy legislation in place, so some issues will be improved as a matter of course, particularly in the commercial sector. But there is a great deal of room for difficulty in government uses of biometrics, where it is much more difficult to see a pathway to meaningful consent. What are the guidelines that can address these issues? We believe the Expert Group can find a way to address this. Ideally, the Guidelines will have global impact. Given this, it is particularly important that the large mandatory biometrics use case problems are addressed, as non-EU countries need guidance regarding commercial as well as government uses. III. Comments on Tension points in AI and Machine LearningArtificial Intelligence and machine learning techniques have matured considerably in the past decade, affording new insights into data across

multiple disciplines. Different flavors of AI exist: Convolution Neural Networks, Markov Models, Ensemble Methods, Deep Learning, Bayesian Belief Nets, Statistical Models. These models have different levels of explainability; there are some interpretable models, some models have the so-called "black box" which can be impenetrable. For some models, modified deep learning techniques can learn explainable features. It is crucial in policy discussions to distinguish between AI models and their differing levels of explainability. Much attention in the past few years has been given to a variety of tension points in AI, for example, the lack of transparency of the "black box." However, additional tension points exist, and should be treated just as thoughtfully. Fairness, transparency, accountability, and good governance around uses of AI and multiple other aspects of AI are among key aspects to include in any principles and policies regarding AI. The Guidance has done an admirable job of incorporating much nuance around these issues. We would like to pause here and in addition to supporting the Guidelines on these topics, also support Japan's AI Guidelines, which in 2018 are now a completed draft after substantive multistakeholder deliberation. The guidelines are thorough, fair, and balanced.We would like to discuss two tension points in particular. That is, input risks, and risks regarding interpretation and use of results. We focus on these two areas here. IV. Comments Regarding inputs/data sets risksAI analysis is a data-intensive discipline, requiring abundant input factors ranging from raw data sets to algorithms, and in some cases, categorizations or scores based initially on raw data sets, a full accounting of the privacy risks associated with input factors is important. First, data sets must be available to use; second, data sets must be appropriately cleaned and prepared for use; and third, the data sets must be appropriately matched to the intended inferences or goals sought from the analysis. These are among the baseline considerations for data sets, understanding that many more considerations exist. Among these considerations includes potential issues relating to data sets that are derived directly from or about individuals or groups of individuals, or in some cases data sets that while not directly derived from or about individuals, can be used to create inferences about individuals or groups of individuals.Regarding algorithms or scores/categorizations used as input factors for AI analysis, a primary consideration (beyond ethical data use and the need for privacy assessments for enhanced risks) is that many of these types of input factors can be proprietary in nature. Given that some AI analysis utilizes numerous algorithms as input factors, proprietary algorithms could pose obstacles for AI use across industries or sectors over time, as well as pose substantial challenges to transparency, fairness, and interpretation. We mention data brokers here as an important category to think about. While data brokers are not as extensively operating in Europe, they are operating in other jurisdictions, and this has impacts on AI inputs and fairness Please see our report, The Scoring of America for many specific details of what this is, how it happens, the products/services in this space, and an analysis and recommendations for

solving the problems. The issue of secondary use of data, and particularly secondary use in AI systems, is important to resolve. V. Comments Regarding interpretation and use of AI outputsHow to interpret the results of AI analysis also needs specific guidance. Interpretation should occur within an understandable, specific context and should be carefully constrained and defined. AI model results are only as predictive or as fair as the score model or models, the factors used in that model, and the training and fit of that model to the task or problem it was meant to solve for, among other factors. However, much interpretive nuance is easily lost when an AI model results in a simple numeric score. A simple score can be deceptively complex to interpret; models can be over or under fit, creating potentially significant discrepancies in results. Over-fitting arises when an algorithm is trained to perform very well on an existing set of data, but has been tailored so well to that data set that it can behave erratically or incorrectly outside of the specific scenario it has trained for. When a predictive model assigns a value or a range to a person, for example, a risk score, the model used to create that value must be transparent, accurate, reliable, and kept up to date. The numeric range for interpreting the result (such as a score) should be well-quantified, and the results validated.Without these protections, even the best and most predictive model can be interpreted improperly, to potentially negative consequences.Currently, very little governance exists around the interpretation and use of specific AI results. It is an area particularly well-suited for further work. Governance models can be used to address the numerous contextual issues that arise in the area of use of AI scores or models. VI. Conclusion The successful design of principles for AI and ML must be effective today, and effective in the future. This is one of the great challenges in composing guidelines on AI and ML today; we are essentially peering into the future and working to anticipate risks in order to mitigate those risks. We offer one idea here: whenever possible, use unambiguous case studies to guide where the harms are. We are certain that India’s Aadhaar system was deeply problematic. We are certain that as other national, large systems use or propose to use biometrics - including DNA — that they need baseline legal and policy protections in place, as well as ethical guidance. We are also certain that gender is an extremely important aspect of discrimination that occurs in AI and ML. There is robust technical support for this, including in the field of biometrics. And finally, after our fieldwork in India, we have come to deeply understand that AI and ML must do no harm and must create a public good. We again voice our strong support the inclusion of these principles and ideas in the draft. Respectfully submitted, S/Pam Dixon Executive DirectorWorld Privacy Forum Notes 1 Pam Dixon, A Failure to Do No Harm: India's Aadhaar biometric ID program and its inability to protect privacy in relation to measures in Europe and the U.S. Springer Nature, Health Technology. DOI 10.1007/s12553-017-0202-6. http://rdcu.be/tsWv. Open Access via Harvard- Based Technology Science: https://techscience.org/a/2017082901/.2 Aadhaar case: Supreme Court of India,

Justice K.S. Puttaswamy (Retd.) and another v. Union of India and others. Writ Petition (Civil) No. 494 of 2012. Decided Sept. 26, 2018. Available at: http://www.worldprivacyforum.org/wp-content/uploads/2018/09/Supreme-Court-Aadhaar-Judgment-26-Sep-2018.pdf.3 There were additional issues related to technical limitations of biometrics, which are well-studied and documented. These technical limitations created harms that the implementers did not anticipate. Across India, government reports faithfully noted extraordinary and mass "failures to authenticate." That is, individuals with Aadhaar IDs could not use their biometric IDs to authenticate themselves. The authentication problems stemmed from failures within the biometric system itself. At scale, statistically low rates of multi-factor or multi-modal biometrics systems can become millions of people who could not get food. In India, there were reports of people dying because of failures to authenticate. Dhananjay Mahapatra, Don't let poor suffer due to lack of infrastructure for authentication of Aadhaar, Times of India, April 24, 2018. https://timesofindia.indiatimes.com/india/dont-let-poor- suffer-due-to-lack-of-aadhaar-tech-sc/articleshow/62842733.cms.4 Kenya’s Registration of Persons Act doesn't specifically mention DNA but has an open list for the data available for collection. In January 2019, President Uhuru Kenyatta signed new amendments into law that changed the requirements for new applicants for National ID cards. See: New ID requirements after Uhuru amends law, January 21, 2019. Pulse Live, https://www.pulselive.co.ke/news/new-id-requirements-after-president-uhuru-kenyatta-amends-the-registration-of-persons/ze50lth.5 Editorial, Address concerns over taking DNA samples from Kenyans, Standard Media, Jan. 29, 2019. https://www.standardmedia.co.ke/article/2001311128/address-concerns-over-taking-dna-samples-from-kenyans.6 Draft AI Utilization Principles 17 July 2018. Japan. The Conference Toward AI Network Society. http://www.soumu.go.jp/main_content/000581310.pdf. These guidelines were crafted with multi-stakeholders and inclusive of Ministry-level experts, academics, and others. The Guidelines were crafted over several years. 7 Pam Dixon & Robert Gellman, The Scoring of America: How secret scores threaten your privacy and your future, World Privacy Forum, April 2014. https://www.worldprivacyforum.org/2014/

Angel	Martin	Johnson & Johnson	There is growing consensus about the revolutionary role that artificial intelligence will bring to society, the economy and the planet. As the leading company in healthcare, we are determined to sustainably embrace those changes looking for new solutions that will help preventing and treating diseases, provide personalised healthcare and give unique patient and consumer experiences. In our view, AI needs to be human centric, and it needs to enhance human interactions not replace them. In order to ensure a prosperous socioeconomic and environmental evolution based on AI, its development and use should respect ethical principles as we currently do, for instance, in health care where bio-ethics is a cornerstone for practitioners, companies and authorities. In our view, we should add Explainability (allowing people to understand users 'if' 'how' and 'why' an AI system suggested a certain decision) to the commonly used bioethical principles: Beneficence, Non-Maleficence, Autonomy and Justice.However, to achieve an ethical, trustworthy and sustainable AI-rich world, we need more than principles. We should also have an appropriate and agile policy framework which: fosters innovation; builds a data culture that enables AI while ensuring personal data privacy protection; supports social cohesion by educating people and upskilling the (healthcare) workforce; promotes ethical behaviour uptake in industry and the public sector; provides incentives for a European AI ecosystem and seeks for international cooperation.	Johnson & Johnson is dedicated to advancing patient care and public health by finding solutions to some of the most complex medical challenges. Our focus on bioethical decision-making stems from the commitment we make to patients, healthcare professionals and customers (1), which is described in Our Credo values (2) and Our Code of Business Conduct (3). An ethical AI world requires the application of ethical principles from the development to the use of AI. AI is like any other new technology; it's value for good or ill is in its application not in the technology itself. J&J is a supporter of AI4People, a forum composed by academics and experts in AI and ethics which have proposed the following principles as the foundation for an ethical AI (4):1) Beneficence (do good): Promoting well-being, preserving dignity, and sustainability2) Non-maleficence (do not harm): Ensuring Privacy, security and "capability caution"3) Autonomy: Ensuring the power to decide (supporting people to make decisions)4) Justice: Promoting prosperity and preserving solidarity (non-discrimination)5) Explainability: Ensuring transparency and accountability (users should know 'if' 'how' and 'why' an AI system suggested a certain outcome over another). While the first four principles are well recognized within the medical community (5), "Explainability" is required to support transparency and accountability in AI. The way in which a model determines its outcomes should map onto the audiences' world model or it will not be comprehensible. This is different for different audiences and so no one approach will work for all applications. It is paramount that the outputs of the algorithms can be properly understood by non-technical audiences, which is necessary to evaluate fairness and gain trust.Johnson & Johnson is committed to ethics-based decision making and agrees that ethics should play an important role in the implementation of AI. Some examples of our work in ethics includes our efforts in compassionate use through the CompAC (6) initiative which ensures fair, objective and ethical evaluations of patient requests for investigational medicine. In addition, we are a leader in initiatives to improve clinical trial data transparency, as evidenced by our commitment to data sharing through the Yale Open Data Access (YODA) Project (7), a model that provides a fair and unbiased approach for assessing external requests for the use of clinical trial data.Furthermore, a complement to Our Credo, Our Ethical Code for the Conduct of Research and Development (8) provides standards of conduct and behaviour for physicians, clinical research scientists and others responsible for medical aspects of research and development. It provides principles that guide ethical decision-making to ensure the safe use of our products, and the best interests of our patients, their families, doctors, nurses and healthcare providers.In order to achieve a fair and trustworthy AI experience in society, ethics should be built into business culture, innovation and practice from the start, providing the necessary means (e.g. education programmes, bioethics committees) to reinforce its application.(1) https://www.jnj.com/office-chief-medical-officer/bioethics-at-johnson-johnson (2) https://www.jnj.com/credo/ (3)	We are committed to partnering with policy-makers and stakeholders to define the foundations for the development, application, ethics and regulation of digital technologies, for which J&J proposes the following recommendations:• Policy approach: Developing a framework that is proportional, risk-based, predictable and innovation-friendly to encompass the evolution of AI technologies and their applications, whereas ethical risk might change drastically according to its use and context. • Data as a key enabler of AI: Encouraging authorities to collaborate with industry and civil society in building data ecosystems which help to generate meaningful datasets in quantity and quality, ensuring and enabling fair and ethical AI ecosystem. Policies should take into account the value of improving citizens' health and healthcare systems through the use of data-driven approaches. These approaches rely on the collection, analysis, and sharing of health data to better understand diseases and treat them as part of a system delivering more personalised 'citizen-centric' healthcare, which is more targeted, effective and efficient (see some examples in the annex). There are still major challenges to overcome, such as data silos, lack of harmonisation, common standards, interoperability, no integration with Electronic Health Records (EHR), lack of adapted regulatory framework and fragmentation of smaller national initiatives. Health care data is one of the most sensitive and needs adequate protection. GDPR is a step in the right direction, as long as its implementation is practical, consistent across geographies and informed of technological evolutions. • Social inclusion and cohesion: Improving educational and professional training systems, for example for digital skills in the healthcare workforce, including in less favoured areas, to make sure that a cohesive and inclusive development and uptake of AI by people takes place across Europe, increasing literacy and improving workforce skills which are essential to understanding and trusting AI. • Self-regulation: Promoting a 'holistic' approach including high-level principles, best practices, voluntary and industry-driven standards (complementing existing regulations). Encouraging industries to self-regulate: companies should establish guiding ethical principles for themselves that will apply throughout all their operations. The process or context in which AI is embedded must also be fair and ethical, which goes beyond the scope of the AI per se and is critical to evaluating the ultimate impact. • Liability: Encouraging AI developers and users to understand the key issues and tools to mitigate risks for end users and patients: As society continue to pilot, adopt and rely on AI technologies to reshape the future of decision making, AI that can be trusted to be transparent, fair, explainable and secure is imperative. • Market access: Supporting an efficient application of existing framework of rules and regulations to validate, authorise and certify AI-based products, for example through bringing AI expertise into regulatory agencies • Privacy -There are aspects of artificial intelligence that are of relevance for privacy. Some systems utilise personal data, while other systems use data that cannot be linked to individuals. If personal data is utilized, appropriate consent must be	• Metrics of trustworthiness: In line with existing professions, such as medical doctors or lawyers, there could be certification of 'ethical AI' to enhance transparency and trust by people on AI applications. Developing agreed-upon metrics for the trustworthiness of AI products and services, to be undertaken either by a new organisation, or by a suitable existing organisation. These metrics would serve as the basis for a system that enables the user-driven benchmarking of all marketed AI offerings. Scientific validation provided by existing regulatory agencies in the relevant sectors (e.g. EMA) also helps to ensure a trustworthy use of AI.In Healthcare there can be multiple use cases. The list below is a short compilation of existing applications of AI:• Innovation: There are AI applications in drug discovery, discovery of new protocols (mining from millions of cases the local protocols or combination of protocols that deliver the best outcomes), or better framing of research hypothesis (today mainly based on intuition and empirical approach). • Mining medical records with Neuro-linguistic programming and Machine learning: huge amounts of information are stored in non-structured files (referral / discharge letters, internal reports, e.g. oncology reports, anatomo-path reports etc). Extracting this information will allow deeper insights and if done "longitudinally", patient disease trajectories can be developed• Designing treatment plans: well curated medical information organized by patient and sorted chronologically, can lead to deeper insight in how patients flow through their disease (ultimately, from health to early diagnosis to treatment and follow-up). While this can be done across all disease areas, there is most activity in oncology, neurology and cardiovascular. This will ultimately lead to better clinical decision support systems• Health assistant & Medication management support for patients: Once deeper insight is derived from steps 1 and 2, it is possible to develop interactive algorithms that help guide patients through their interaction with the healthcare system; so called virtual nurses. A specific application of this are chat-bots (later voice-bots) – an important early example is Babylon Health in the UK – but now also collaborating with the largest social media provider from China (Tencent and WeChat). • Diagnostics – wearables and sensors: traditional (e.g. in ICU) and novel diagnostic tools (e.g. wearables and various sensors) create a tremendous amount of data which needs to be analysed "in context" of the overall health or disease state and of the individual's constitution (phenotypical and genotypical). The ability to combine many more data points on the person could open the path to identification of pre-conditions and early diagnosis.• Image analysis and assisting in repetitive jobs: AI is extremely well positioned for pattern recognition in "digital streams" – images offer such "stratum". Before effective analysis is possible, images must be annotated (by humans) so that the pattern recognition algorithms can be trained. Multiple examples are available, mainly in ophthalmology and oncology. Belgian companies such as Robovision (https://robovision.be/) are good examples • Precision / Personalised medicine: combination of many of the above-mentioned elements will lead to a very	Artificial intelligence is bringing a new revolution to society, enhancing people's interactions and capabilities, and helping to improve and personalise goods and services and thus resulting in unprecedented patient and consumer experience. In healthcare, AI has the potential to help address some of the biggest challenges in certain therapeutic areas as well as in healthcare systems in general with more efficiency and better outcomes across the patient pathway. AI impacts the entire value chain from R&D and clinical trials to supply chain giving faster access to better and personalised drugs and treatments for patients. AI will play a key role in preventing diseases, providing better diagnosis and even helping doctors to dedicate their time to higher-value activities and be freed up from admin activities (medical record keeping). It will also dramatically change self-care and democratize access to personalised treatments and consumer products.At Johnson & Johnson, we are committed to designing people-centric solutions which can help diagnose, intercept and treat diseases early, and empower doctors and surgeons leading to better and more personalized care. J&J is currently applying Machine Learning for the discovery of new drugs, the optimisation and personalisation of surgical instruments and implants, for example in creating algorithms to optimize the use of the device in surgery based on the tissue type, as well as in the identification of rare adverse events and pharmacovigilance. Since 2015, Janssen Research & Development has three new research platforms focusing on disease prevention, disease interception and the microbiome – areas of transformational medical innovation that are expected to change the healthcare landscape. Disease interception will intervene earlier than today's clinically accepted point of diagnosis and seek solutions that stop, reverse or inhibit progression to that disease, for instance type 1 diabetes or various forms of cancer. We are also developing solutions that effectively respond to consumers' search for personalized products that can meet their needs. For example, our latest innovation, the Neutrogena MaskiD™, is supported by AI to personalise skin care, using 3-D cameras and advanced 3-D printing technology. AI is also helping us to identify consumers and stakeholders' insights from unstructured data to find unmet needs and improve both products and communications. Furthermore, we are using AI assistants to retrieve and organize information enabling more efficiency in our operations, as well as AI-powered capabilities from leading digital media partners.

				https://www.jnj.com/_document?id=00000159-69fe-dba3-afdb-79ffcdd60000 (4) http://www.eismd.eu/wp-content/uploads/2018/11/Ethical-Framework-for-a-good-AI-Society.pdf (5) Nuremberg Code (1947), Declaration of Helsinki (1964), Belmont Report (1978), The Common Rule (1991)(6) https://www.janssen.com/compassionate-use-pre-approval-access (7) http://yoda.yale.edu/johnson-johnson (8) https://jnj.brightspotcdn.com/2e/33/ec4721ff45f19b4a07137a17a607/ethical-code-for-the-conduct-of-rd.pdf	obtained, and use must be in line with the purpose for which it has been collected. • Explainability: Developing a framework to enhance the explicability of AI systems which support the decision-making on socially significant issues, such as healthcare. In order to build societal trust, it is important that the workings of the models can be explained to non-experts. Therefore, central to this framework is the ability for individuals to obtain a factual, direct, and clear explanation of the decision-making process, especially in the event of unwanted consequences. This is likely to require the development of frameworks which are specific to different industries.• Incentivise development and use in EU: Incentivising (financially) the development and use of AI technologies within the EU that are socially preferable (not merely acceptable), and to develop an AI ecosystem in healthcare, including public funds for infrastructures and tax incentives.• International cooperation: Encouraging policy-makers to cooperate at international level on ethical guidelines, helping to ensure an inclusive and global approach.	comprehensive patient record in the context of an overall population. This will provide insights into sub-segments of patients which present similar responses to interventions driven by their specific "make-up". AI will boost usage of data along the entire patient life cycle, from early detection – treatment – outcome prediction• Precision surgery: AI will open the door to decision support during surgery based on factors such as patient conditions, instruments sensors or critical structure recognition, helping to deliver shorter post-surgery hospitalizations and better patient outcomes• Health system analysis and optimisations: Pattern recognition analysis on prescription and reimbursement data would generate a more efficient healthcare support model, for payors, governments, industry and patients, moving towards a value-based service models where the focus shifts to reimbursement based on patient outcomes. It should also allow for optimization of administrative and work flow task, but also to identification of possible mistakes (e.g. medical and/or prescription mistakes) and fraud detection (e.g. falsified prescriptions).	
Ana	Garcia Robles	Big Data Value Association	The AI opportunities (e.g. to improve our environment, to create new business opportunities etc.) of course following ethical guidelines (usable also in adaption for different sectors and from different companies/SMEs) is missing. Not all AI applications/solutions will have the same impact/influence in humans and therefore are equally linked to an Ethical purpose. Tailoring is needed per AI application.	The specific recommendations and guidelines should be more aligned to existing processes for security, data protection, product safety and security. Should build on more on the existing concepts like GDPR and for example not limit data processing for AI exclusively on the basis of informed consent.	Comments to the "Requirements of Trustworthy AI": - Accountability: In general, the accountability should be at least as strong as the accountability of a human for the same activity. - Data Governance: Some recent market failures (e.g Microsoft chat bot, Tay) demonstrate that we cannot always rely on self-correction due to large enough data. - "Design for all" or "one-size fits all" concept is not applicable to data driven AI applications (e.g. if an AI application is made for a specific purpose). BDVA can help with the expertise of our member from different sectors to define this more clearly. Comments on the "non-technical methods": - All AI systems should come with a clear description of their limits, including the areas they are intended for and those, they are not intended for.	- BDVA can help to define and implement necessary guidelines which can also be used from SME and helping to improve European economy with the help of AI. - BDVA recommends to do proof of concepts of the guidelines (e.g. with help of our i-Spaces) before guidelines will be published. - Special attention is needed to the assessment criteria to ensure that we strike a good balance and not hinder innovation and growth of the EU industry in the field of AI.	AI definition should be improved. BDVA can help to provide this definition based on the know-how our 200 members. Fostering R&I on achieving Trustworthy AI in EU should get a prominent position in the document, at the forefront of activities, including practical test cases (e.g. i-Spaces) for various verticals. Industrial AI is missing in the entire document. BDVA has a positioning paper defining data driven AI which could be used and we believe that data driven AI as a key differentiating strength for EU of macro-economic importance. BDVA also offers the help to support with all these topics. Responsible Research and Innovation is completely ignored in the document. References to the work developed in the last few years in this policy area could enrich the content of this Ethics guidelines.
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	

William	McSweeney	The Law Society, Technology and Law Committee	The Law Society welcomes the development of standards and best practices for responsible development and use of AI. We believe such standards and best practices should be developed with the participation of all relevant stakeholders and be subject to open and thorough debate. Although the European Union is certainly well positioned to take a leading role in that work and serve as a "benchmark" for other countries and international organisations developing similar standards and best practices, the very nature of AI development and use in the modern interconnected world dictates that such work, to the extent possible, should include and encourage the participation of all relevant stakeholders irrespective of their geographical location and international association.	The Ethical Purpose and Use of Trustworthy AI:The starting point of the report to include "Ethical purpose" as one of the two fundamental components of a "trustworthy AI". We believe, in so far as it concerns the development of AI, this requirement creates serious practical challenges, as it implies that AI developers must find right answers to all kind of ethical dilemmas and always adhere to values of "a democratic society", which is not realistic to expect, as it would be impractical to suggest that a computer (or for that matter any piece of technology) must have "ethical purpose", i.e. can only be developed, deployed or used if it advances the "good life of individuals".To have a true assessment of ethical purpose, one must look at how the AI is used, applied and implemented in different contexts and sectors. Therefore, we would suggest adding a definition of 'Ethical Use' to the guidelines. Such a definition would draw on the principles set out in the report, i.e. protecting fundamental rights and democratic principles in applications and in system interactions. The definition should also include 'explicability' as described on pg. ii.Balancing Ethical Values: When there is no expert consensus on ethical values and potential long-term consequences of 'untrustworthy AI', how is it possible to balance ethical values in the design, evaluation, development, dissemination and deployment of these technologies?"Critical concerns raised by AI" states "the inability of the AI HLEG members to agree on the extent to which the areas as formulated raise concerns." The "dual-use nature" implies the inability to agree on balancing different ethical values and the lack of clarity raising the likelihood of long-term unethical use. Guidance Applicability: In addition, some ethical values are specific to different cultural contexts and sectors. How can we equate an AI system developed targeted at vulnerable consumers, with a Business to Business solution? A government body using an AI system for the public, with a private commercial enterprise using an AI solution for (i) internal use or (ii) product or service offering to the public? The debate around the 'Ethical Purpose' of AI is made less useful when ethics are applied uniformly to all contexts.	Non- Technical Methods:We would propose that it is only through the use and implementation of the AI that the ethical values and rights can be assessed and that be a continuous assessment. In the non-technical methods to achieve trustworthy AI (p22) the following are highlighted: - "Make Trustworthy AI part of the organisation's culture"- "Ensure participation and inclusion of stakeholders"- "Ensure a specific process for accountability governance" All organisations should have these principles of accountability, codes of conduct, data governance and need to involve key stakeholders. These should be elements of corporate responsibility to employees, the public and to consumers of their products. However again, not one-size fits all. There is more work to be done on the effectiveness of an organisation's culture and accountability governance in the context of AI. The report needs to address more explicitly how organisations are making decisions on whether or not to adopt AI and build this into the existing rules around corporate governance (which may be covered in the second report)."Organisation structure and capacity for adopting AI"One challenge that is not fully addressed in the Non-Technical Methods Section of the report (pp 21 & 22) is the ability of organisations to have effective structures in place to adopt AI. How organisations are structured can add difficulties to adopting AI and limit capacity to consider ethical values/rights (whether for example the AI is for internal business functions or to enhance and deliver a product/service).For some organisations (with the exception of high tech companies), the decision (on whether or not to adopt AI) is in its infancy - not on the radar of the senior members of the organisation/company boards or AI is restricted - as it is viewed as a distinct, self-contained area, separate from the rest of the running of the organisation. An area which some would say, speaks a different language. As an example, the product teams, who may be looking at AI, are usually located separately from the central corporate, finance, sales and consumer teams. AI is therefore seen in some organisations as an "add on" and hence capacity is not allocated. "Capacity" can include resources, structure and of course, funding. If the ethical considerations highlighted in the report and as we recommend Ethical Use, are to be applied in the business context, consideration needs to be given as to how to integrate AI into the fabric and running of the organisation, so that it becomes one of the organisation's core responsibilities. Data Lifecycle: The draft should illustrate and educate how ethical considerations, and bias, can be analysed at each stage of the proposed data lifecycle (Pg. 19). This would provide entrenched technologies, and their creators, the chance to examine how changes could be made to their systems without placing their security and users vulnerable.	We think that the proposed assessment list largely reflects the necessary aspects of developing the AI systems. In addition, we suggest adding 'Responsibility' as a first step in the assessment list (i.e. before "Assessing Trustworthy AI"). Responsibility would refer to how the decision to adopt the AI is made within the organisation and would address such questions as:- Who is responsible?- Who are the stakeholders to both input and sign off on the recommendations for adopting particular AI system?- Are they clear on the requirements/specification for the AI?- Is the AI being produced in-house, bought off the shelf, open source, or specially commissioned?Since there are different routes to adopting AI, the decision on this route will be relevant to other items on the assessment list, including liability, accountability and control of remedying any issues (part of Governing AI autonomy on the list). Transparency:The Transparency question 10 (page 27) does touch on some of the points above, and of course there are the detailed points to be addressed, how the system works in terms of data protection, and security for example which would be included in the spec, but operationally all businesses (including law firms) need to have the due diligence and assessment of the first step.	If the Guidelines were more focused on the issues of technical and organisational measures of mitigating the risk of "unethical use" of AI, they would be more helpful in constructing the future debate around "trustworthy AI. We would advise developing a set of principles we mention in Chapter II, and perhaps suggesting practical examples of such technical and organisational measures being applied in practice, some of which might be "borrowed" from other fields of research and development facing ethical dilemmas (genetic engineering, medical research, etc.).Trustworthy AI and ethical use of AI is heavily predicated on trustworthy and ethical use of data in general. So, setting the "ethical purpose" as a fundamental principle for AI development, implementation and use without having similar standards on the use of data in general might lead to a perverse situation where higher standards of ethical behaviour are demanded of machines than of humans.Finally, having "ethical purpose", as defined in the Guidelines, as one of the two fundamental components of "trustworthy AI", which we consider impractical for the reasons explained above, creates the danger that the future debate around ethical AI might become too theoretical and part ways with real-life developments in this area.

Thorsten / Joachim	Gantevoort / Iden	TÜV Rheinland	Rationale, page 2 “The role of ethics”comment: in later chapters it will become more evident that a discourse on ethics of AI will lead to the necessity of considering meta-ethical aspects. Which ethical principles shall form the basis of the discourse? (deontological, requirement based (ref. I. Kant), utilitarian (ref. Bentham), virtue-based, ….); There is also the need for professionally trained specialists. Philosophy, in form of ethics is to become part of the engineering process.	1. The EU’s Rights Based Approach to AI Ethicsp 5 “The field of ethics is also aimed at protecting individual rights and freedoms, while maximizing wellbeing and the common good”comment: this statement is probably philosophically not viable and should be replaced with a statement like “the proposed approach is also aimed at ….”4. Ethical principles, p 8, paragraph 3 “…. in particular situations, tensions may arise between the principles ….”comment: this underlines the above comment regarding the need for a meta-ethical discourse, which must be part of all stages of design, deployment and assessment of AI systems.4. Ethical principles, p 10, paragraph 4 the principle of explicability, operate transparentlycomment 1: there are many relevant interpretations of “transparency”: a) transparency as traceability in design and testing, b) interactive transparency, allowing the user to correctly anticipate imminent actions of an AI system, c) state-space transparency (w.r.t. online monitoring), d) forensic transparency (‘black boxes’), e) ‘psycho-moral transparency’ as not being deceptive about the nature of the AI system, which is not a sentient being (ref. J. Bryson, “patiency is not a virtue”), f) transparency with respect to costs and side-effects, absence of hidden costs or side effects, g) explicability/explainability….All of these interpretations need to be taken into accountcomment 2: “explicability” must have three components:- intelligibility- truthfulness- completeness“completeness” presently does not appear to be explicitly included.Further, “relevance” can be seen as one aspect of “truthfulness” but it is important enough to be mentioned here separately -- an explanation must apply to the matter that a user expects. From this also follows the need to update the user about any changes of conditions and characteristics of a system or service in the future. “Explicability” shall not be misusable to formally obtain consent at a single point in time in order to conclude a contract, but must be the foundation of the relationships between user and system and user and service provider on a continuous basis.comment 3: The concept and level of explicability need to be discussed. It is unlikely that (all) users will have the same level of ability to understand the functioning/possible reactions of AI or even might “identify” AI. The question is therefore, who shall be able to understand (i.e. to assess, to supervise …) AI?Implementation of this principle shall not be used as an argument to release the manufacturer, operator, government, etc. from the responsibility to apply protections and to establish monitoring/governance for the AI itself as well as for the designers and operators.5. Critical concerns, 5.2 p 11 Covert AI Systemscomment: There are at least two possible implications here: either the development and deployment of systems which could be confused with humans must be strictly regulated or such systems must be legally protected against vandalism. Also, see comment below regarding “potential longer term concerns”. The paragraph may also be understood to contain some arguments against nudging.5. Critical concerns, 5.4 p 12 Lethal Autonomous Weapons Systemscomment: the guideline shall oppose development of lethal autonomous weapons, as the primary	II. Realising Trustworthy AI, 4. Governance of AI Autonomy/Human Oversight, p 16 Governance of AI Autonomy/Human Oversightcomment: the sentence “this also includes the predicament, that a user of an AI system [….] is allowed to deviate from the path [….] recommended by the AI system” is either misleading or contains a mistake. The formulation in the text, appears to imply that human autonomy amounts to a “predicament” which should be minimized.II. Realising Trustworthy AI, 8. Robustness, p 17 “Accuracy”comment: “Accuracy” relates to machine learning specifically. Not to other technologies. Further, as “reproducibility” is mentioned as critical requirement, more explanation is needed regarding the difference between reproducibility and determinism.2. Technical and Non-Technical Methods, p 19, paragraph 4 “This also entails responsibility for companies to identify from the very beginning the ethical impact that an AI system can have, ….”comment: the ethical analog of a HAZOP needs to be performed, requiring both experts to perform such analyses and specialists who train these experts.2. Technical and Non-Technical Methods, p 19, paragraph 5 Architectures for trustworthy AI: “This can either be accomplished by [….] and the monitoring of which is a separate process.”comment1: this is basically the idea of the “ethical governor” (Arkin et. al., 2012). Such an approach makes specifically sense when the “ethical governor” is of lesser complexity than the system to be monitored, because it can be validated more completely. It is however, not clear that this will be achievable for all relevant systems, as the ethical governor itself must have the capacity to judge the context of an action.comment2: systems can fail even when an “ethical governor” has been successfully validated and implemented. Further design-specific measures must therefore be considered to reduce the possibility of catastrophic failure at a high level of user trust. (A system must be considered ‘opaque’ to the extent that serious failures may occur at high levels of user trust)2. Technical and Non-Technical Methods, p 20 Testing & Validationcomment1: the difficulty will be to show, for any combination of techniques, that this combination is sufficient in a well-defined, objective sense.comment2: the sentence “Intelligence manifests itself on the semantic level, e.g. during program execution” is neither helpful nor required here. It is enough to refer to complexity, lack of reproducibility and the inexhaustibility of naturally occurring input combinations.2. Technical and Non-Technical Methods, p 21 “Regulation”comment: Legislation may increase trust, but trustworthiness depends on the actual implementation. Legislation/regulation only increases the likelihood that trustworthy systems are being built, not the level of trustworthiness of those systems.2. Technical and Non-Technical Methods, p 22 Education and awareness to foster an ethical mind setcomment: again this points to the need of having professionally trained ethics experts, who are aware of the necessity of a meta-ethical discourse	III. Assessing Trustworthy AI, p 24comment1: suggested is here to establish the notion of and protocol for execution of an ethics-oriented HAZOP, addressing- what is the purpose of a proposed system?- what are its intended benefits?- what is the worst case scenario in case of system failure?- what is the worst case scenario absent of system failure?• what are its overt costs and side effects?- how can the absence of covert costs and side effects be confirmed?When costs and side effects are determined, consideration shall be given to - 1st person risks (risks to the user)- 2nd person risks (risks to persons intentionally participating in the user’s activity)- 3rd person risks (risks to persons randomly encountered and not intentionally involved)- nth person risks (risks to persons in other locations, who are usually not encountered by and whose existence may even be unknown to the user)- environmental risksPerformance of such an analysis will require expertise in diverse fields: ergonomics, psychology, social sciences, economics, environmental sciences, ….comment 2: the causes of detrimental effects of a system are not limited to system failure, but include lack of ethical alignment of their purpose and unintended and unforeseen psychological effects (e.g. ‘smartphone addiction’)comment 3: for the assessment list, also consider: system override, system decommissioning, fall-back plancomment 4: in assessing risk levels and their indeterminacy, priority shall be given to quantitative assessments. When quantification is analytically not possible, other methods need to be devised. One possibility may be to combine the rated opinions of a number of experts according to well-defined semi-numeric procedures. In any case however, all findings need to be documented and not discarded, even if quantification proves to be impossible. Such risks shall be categorized as “not (analytically) quantifiable”. Records of identified non quantifiable risks shall be reviewed periodically.III. Assessing Trustworthy AI, p 26Respect for human autonomy / “[….] risks on human mental integrity (nudging) by the product [….]”comment: nudging is in conflict with some aspects of transparency. Deployment scenarios must be strictly regulated and qualified and sufficient human oversight must be assured.III. Assessing Trustworthy AI, p 27Fall-back plancomment: it is important that this item is mentioned here. In addition, it must be contemplated:- what are the actual levels of risk and societal impact?- how can the uncertainties about the involved risks (severity and probability) be assessed?- what are the economic costs of implementing viable fall-back plans?- is society willing to carry the costs for deploying and maintaining viable fall-back systems or infrastructure?- high impact potential at high risk uncertainty should lead to rejection of a technology	Please be aware that the elaboration of these comments was performed by Joachim Iden and myself (Dr.Thorsten Gantevoort), see "double names" in contact sheet.Executive summary, p I, 2nd paragr. “advance in AI techniques, such as machine learning…”comment: that depends on the definition of “AI”; a notion like “autonomous and/or artificially intelligent systems” may be more encompassingConclusionp 29 “The HLEG recognizes the enormous positive impact [….]”comment: cite evidence or delete. This comment is not needed. It would be sufficient to state that there is a positive impact potential. On the other hand, there is also the possibility of detrimental impacts and this is where the present guideline needs to contribute to methodologies and measures to avoid and mitigate the corresponding risks.Conclusionp 29 “Trustworthy AI [….] has been our north star”comment: a more neutral expression like “guiding principle” would be preferred unless the purpose is to make the point that the suggested framework is only applicable to humans and systems located in the northern hemisphereExecutive summary, p I, 3rd paragr. “AI’s benefits outweigh its risks…”comment: it is not obvious that this is the case; the purpose of the guideline is exactly to ensure that benefits will outweigh the risks, but this will also depend on the methodology and details of ethically oriented impact and cost-benefit analysesExecutive summary, p I, 3rd paragr. “human beings will only be able to confidently and fully reap the benefits of AI if they can trust the technology”comment: there are a number of factors of societal, psychological and material nature, influencing the degree of utilization of a technology and the level of trust, potentially resulting in underutilization, overutilization, undertrust and overtrust. Trust is neither a necessary nor sufficient condition for utilization. Similarly, trustworthiness is neither a necessary nor sufficient condition for trust.References• Arkin, R., Ulam, P., Wagner, A.: Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception, Proc. IEEE, 100 (3) (2012), pp. 571–589• Bryson, J.. Patiency Is Not a Virtue: AI and the Design of Ethical Systems. AAAI Spring Symposium Series, North America, mar. 2016

				purpose of such systems is in conflict with the basic ethical principles; the continued study of the impact of such systems however, shall be supported5. Critical concerns, 5.5 pp 12/13 potential longer term concernscomment 1: regarding rights for AI systems it may be recommendable to employ a pragmatic and cautions approach: at least certain systems with a humanoid or zoomorphic component should be protected from vandalism for the same reason that I. Kant advocated protection of animals from human cruelty: protection of humans against a potential worsening of the social climate and resulting in increased levels of violence. Such systems should be treated as 'quasi moral patients'. Such principles, if established, could eventually be extended to treat certain systems as actual moral patients and agents, should such need arise.comment 2: another consideration should be added: the necessity of continuous reflection about the worst case scenario in employing AI technologies.	
Laurent	ZIBELL	industriAll European trade union	Artificial intelligence (AI) is already in use all over the EU, even if it is invisible.In the form of "automated decision-making", the workings of AI are often deliberately opaque in order to protect – open and hidden – corporate interests, for instance in 'social scoring', credit lines, social bots, nudging. AI is not just about technology or software programs, but societal choices are incorporated in this automated decision-making. A debate about discrimination, equality, social justice, participation in relation to AI is needed. It should be clear that AI should not discriminate, it should strengthen equality, enhance social justice and participation. In the form of automated driving of industrial machines and processes, or of vehicles, AI embodies the competences of skilled industrial workers, as well as the information contained in a wealth of data automatically generated by these machines, processes or vehicles. Digital monopolies currently are in a process of private appropriation of these skills and data, and of concentration of the wealth thus generated. This is a major distributional problem, as well as a major risk for the livelihoods of skilled industrial workers.What is needed: sustainable AI made in Europe - ecological, fair, inclusive.Such a comprehensive approach can't be limited to ethics. The debate needs contributions from sociology, philosophy, political science, industrial technology, economics and data experts. The focus of the discussion must be on the politically relevant questions – at national and at EU-level. IndustriAll Europe welcomes the approach to connect AI with European values and principles. This is a first step in the right direction, but more steps are needed. New technology, and in particular Artificial Intelligence , must be shaped in	This legislation regulating Artificial Intelligence should bear upon the following aspects, in addition to those already described in the document:* the usage of AI to supervise work and to profile workers should be regulated, and allowed only after the collective representation of workers at the right scale (i.e. trade unions, or works councils where relevant) have consented; * human workers must be able to take decisions different from the "recommendation" made by the AI system, and yet not be sanctioned for having done so when this decision proves to be wrong;* human workers must be able to test, experiment and innovate, even against the "recommendation" made by the AI system, and yet not be sanctioned for having done so when the test / experiment / innovation fails;* AI systems must be sufficiently reliable and their behaviour must be reproducible enough to ensure safety of material systems (specifically: of machines in a working environment), and particularly of "safety critical" systems where failure is known to cause deaths in large numbers (e.g. civil aviation, rail equipment, chemical plants, civil nuclear power);* AI systems must only be deployed in safety-critical applications after the level of explicability of the decisions, and the capacity to trace back an accident or incident to its cause, are sufficient for this cause to be treated, and for the safety of the application to improve over time; workers must be trained to deal with AI in particular to apply the emergency brake where necessary;* robots (aka "chatbots") must be identified and visibly marked in all on-line debates and discussions, so as not to be mistaken with genuine human opinions, or even be prohibited from taking part in some on-line	The requirement of "distributional fairness" must be added to the list of "Requirements of trustworthy AI" given in §II.1."3bis: Distributional fairness:The added value created by AI must be distributed fairly in society and economy, specifically by making sure that the access to the data that teaches AI systems is broadly distributed among all economic players under Fair, Reasonable and Non-Discriminatory (FRAND) legal and economic conditions, and cannot be captured by digital monopolists."	

way to avoid a threat to democracy and functioning markets, and to avoid further concentration of wealth and power in the hands of very few digital monopolists. First and foremost, it has to be determined which of the challenges posed by AI can be addressed by enforceable rules and laws and which can be left to unenforceable ethic codes, guidelines, self-regulation or voluntary self-commitments. In modern democracies it must be a principle that its three cornerstones, (1) the principles of democracy, (2) the rule of law and (3) human rights, must from the outset by design be incorporated in AI.Citizens and workers, in particular workers' representatives in industrial companies must be empowered to understand the new challenges ahead and be enabled to find appropriate answers. The GDPR was a first step in the right direction, but more regulation is clearly needed (for self-improving industrial machines, AI-assisted maintenance and repair, self-driving vehicles, face recognition, drones etc.) . The Commission should play a role to launch such a holistic debate involving a wide range of stakeholders and contribute to close the gap between Member States.The focus of industriAll Europe lies in the world of industrial work, in particular the future of industrial work. AI needs to be embedded in decent work. AI is ambiguous and needs to be shaped, it can be used to cement power asymmetries or to dismantle them. It is in the interest of workers that information, consultation and board-level participation rights as well as collective bargaining are respected and fully applicable. A general information of stakeholders is clearly insufficient. The rights to information, consultation and board-level representation must cover the area of AI. A technological and social impact assessment is necessary as well as participative research to follow the design, application and implementation of AI and its economic and social consequences.It is of utmost importance that enforceable regulation creates an appropriate framework for AI in Europe. IndustriAll Europe subscribes to a 'human-in-command'-approach to AI so that final decisions are taken by human beings and not algorithms. AI as digitalization in general has the potential to liberate work from dangerous, monotonous and repetitive tasks, in the same time allowing surveillance and control in a totally new dimension. In order to harvest the potential and to minimise risks, it is necessary that trade unions and workers' representatives in general, and in company boardrooms in particular, regularly scrutinise and closely monitor the introduction of new technologies and AI. In particular it is important to ensure that AI fits with the targets of EU climate, energy and environment policies. AI cannot work in a lawless zone where chatbots are not identifiable, can contribute to hate speech, influence democratic elections and undermine democracy itself. In view of the upcoming European elections, but also in democratic discourse generally, it is important to know whether one's counterpart is a human or a machine, which is not the case currently.The rules for AI are not yet in place and it is important to take the necessary steps.The respect for human rights, for workers' rights, for humans' moral and physical integrity, and the cohesiveness discussions (e.g. on political, social or moral issues, in particular during election campaigns);* the added value created by AI must be distributed fairly in society and economy, specifically by making sure that the access to the data that teaches AI systems is broadly distributed among all economic players under Fair, Reasonable and Non-Discriminatory (FRAND) legal and economic conditions, and cannot be captured by digital monopolists.

			of our societies are fundamental goals, which cannot, and should not, be left to the free appreciation of businesses regarding their marketing or communication strategy. The reaction of the EU to the very real threats posed by AI to the achievement of these goals cannot restrict itself to indicative guidelines, with no external scrutiny and no sanction in case of non-compliance. IndustriAll Europe thus demands strong, enforceable, regulation of AI, based on legislation. EU-wide legislation has the advantage of preventing downward regulatory competition among Member States.The legislation should prescribe procedural steps and institutions within organisations to ensure the trustworthiness of AI applications (under the model set by the GDPR), which can be verified by any layperson, and should limit to a maximum "ethics panels" or "boards", often self-serving, which do not provide sufficient predictability of their decisions and/or are vulnerable to conflicts of interest.				
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Carlos	Rodriguez Cocina	Telefonica	• “Trust in the business model” is identified in the introduction as one of the three pillars for a trustworthy AI, but only referenced again in the Principle of Explicability (Transparency). Business models enabled by AI technology should not pursue an unethical purpose and thus be included within the Transparency requirement to realize a trustworthy AI (Section II) and also reflected by example questions in the Assessment List both on Transparency and Fairness (Section III). • It is much welcomed that the Scope of the Guidelines acknowledges that different situations raise different challenges by referring to concrete examples of AI systems: recommendation of songs and of critical medical treatment. Along these lines, it should also be acknowledged that based on such different challenges raised by different situations, guidelines and related obligations should be graded, or applied with different intensity according to the impact a specific AI based system has throughout all the levels of AI system life cycle (development, deployment and usage). The same way cybersecurity requirements are different for a domestic watering IoT device vs nationwide energy grid, AI principles should apply differently depending on its impact in order not to inhibit innovation of simpler, lower impact AI based systems. Therefore, it should be emphasized that requirements for trustworthy AI (Section II.1) and, even more so, the technical and non-technical methods to achieve trustworthy AI (Section II.2) should be domain and application specific.• What is the evidence that fostering a human-centric approach to European AI will enable Europe to become a globally leading innovator in ethical, secure and cutting-edge AI? Current leading institutes for ethical AI are non-EU based, e.g. in NY (AI Now), the Singapore government has already created an AI ethics commission, and the UK is also setting up initiatives (e.g. Ada Lovelace institute). In fact, Europe should act, as other	• On the principle of Beneficence, “Do good”. We do not discuss the importance of doing good as a great ethical principle. What we challenge is the opportunity and applicability to AI (or to any other technology). Even more, in principle, rules (and these are rules) should not demand others (or the technology) to do good but to do no harm. Boldly applying the “Do Good” principle would restrict companies’ freedom to innovate in, or perform regular businesses to the extent that their primary objective might not be improving collective wellbeing. A relevant case among many would be the use of AI in advertising, which some might argue is not aligned with the beneficence principle. This would limit Europe’s opportunity of learning to use AI through marketing, particularly advertising, while this activity is a low risk / highly profitable form of AI (as compared to other AI based decisions with greater societal impact and thus risk, such as a healthcare decision) which could in turn enable the funding of more AI research in the EU. Considering Europe is lagging behind other regions in the use of data for marketing purposes, restricting the use of AI for marketing purposes based on the “do good” principle would have the opposite desired effect of this guidelines. The “do good” principle should be modified in order to provide room for these activities (such as advertising). • The “Do not harm principle” states negative profiling should be avoided. While it is already clear that the do not harm principle enshrines eliminating all negative actions, “profiling” is neutral from a normative perspective, and what makes it harmful is the purpose of the profiling, which relates to the business or public governance model. In fact, profiling is widely used and needed for whatever commercial activity. At the end of the day, we are profiled countless times by every digital interaction we have, with or without use of AI. Thus, unless it is clearly explained what is to be interpreted as “negative profiling”, we would ask the	• Data Governance should add labeling of data as a best practice in order to assure accountability, explainability and improvement of AI training and validations tests, and thus be also able to assess the quality of the data itself. • Data Governance is a known term in the area of Big Data, and has a broader meaning than the intended in this section. We suggest to change to Data requisites. One of the main aspects mentioned here is about bias in data sets, and the importance of being aware of this and correcting it. The section should include a reference to ethics around data gathering or the use of AI to coerce data collection• Accountability Governance as a form of non-technical method should include a reference to auto-regulation, self-regulation and the procedures through which the Governance framework assesses compliance.• The three dimensions for trust in AI (technology, data governance and business model) should be the first bullet in the summary box KEY GUIDANCE FOR REALISING TRUSTWORTHY AI.• First phrase of “Non-Discrimination” needs editing: “Discrimination concerns the variability of AI results, between individuals or groups of people based on the exploitation of differences in their characteristics (such as ethnicity, gender, sexual orientation or age), that can be considered either intentionally or unintentionally, which may negatively impact such individuals or groups.”• Monopolization of data is a critical, immediate threat to the development and implementation of AI and could be an ethical question in and of itself. Since access to data, and more relevantly behavioral data, is a requirement for the development of AI, and mostly all these data is going to be in the hands of a few companies, AI will be controlled by such companies. Indeed, the ability to track the behavior of billions of users worldwide, through the provision of multiple conglomerate services, is possible just and only for an extremely short list of global	Recommend the addition of the following question within the assessment list:• Respect for privacy:o Are end users informed of which data sets are being treated, for which purposes and which is the expected output of those treatments?	Overall the Guidelines provide a good and thorough approach to ensure that AI will have much more good use than bad (intentional or unintentional) use. We agree with that, evidenced by the fact that, in October 2018, Telefonica has voluntarily published its Company AI principles to foster a trustworthy environment for our stakeholders regarding how we will develop and use AI (https://www.telefonica.com/en/web/responsible-business/our-commitments/ai-principles), and by Telefónica’s more general Business principles (https://www.telefonica.com/documents/153952/388559/OurBusinessPrinciples.pdf/adfea195-d91a-4718-8c6f-760f07f4cbdb) where transparency and acting in accordance with non-negotiable ethical standards are two core principles.In this “general comments” part, we distinguish our feedback in two parts: 1) key concerns on the objectives and scope of the draft guidelines, and 2) general feedback on the content of the draft guidelines.1) Key concerns on the objectives and scope of the Guidelines.• We think it is important to explicitly state that the stakeholders invited to voluntary endorse the Guidelines should not only include European organizations, but all organizations that serve EU citizens, businesses and governments, wherever in the world they are based.• It is difficult to assess how these Guidelines will increase the competitiveness of Europe without having the opportunity to analyze how the AI Policy & Investment Recommendations promotes the development of AI capabilities, which is key for understanding how Europe plans to catch up with other regions around the world. • In the same line, we are concerned about these voluntary Guidelines turning into Regulation, especially if that new Regulation would only apply to, or be enforced on, European businesses, and not to businesses in other regions of the world that are serving European customers. Since most of

countries/regions could put in practice an ethical AI approach much earlier. Europe's advantage lays in its ability to drive a collective agreement and act as a block. This will be valuable in the future as the world grapples with the cross-border challenges of AI.

deletion of the reference to profiling in the "do not harm" principle.• On "informed consent" as a value. The Guidelines consider informed consent as an ethical value, which puts in practice the fundamental right of human dignity (sic) and makes a direct link with explicability. This does not take into account that even the GDPR considers five legal bases for processing other than consent which of course do not negatively impact human dignity. Additionally, transparency (explicability) should not be inextricably linked to consent but applied independently of the concrete ground for the processing. As such, we would recommend including a different example on how to go from fundamental right to principles and values. • It is questionable that Covert AI systems by themselves represent a critical concern; it will depend upon the function the system provides. If for example AI system is used for speech recognition for a more advanced IVR system, it is not a really a concern. In fact, IVR does not announce itself as a machine and we don't question this now, should it? Does it need to since it is not AI? But is a synthetic voice AI? In this case, this is more a transparency related issue than a critical concern.• Providing examples on Potential longer-term concerns, such as Artificial Consciousness, given that there is no accepted or consensus theory around the topic, just serves to increase unfounded concerns. As the aim of the paper is to provide the foundation for trustworthy AI, giving such futuristic science fiction like vision on AI just serves the opposite objective, raising alarms without providing any solution or mitigating effect. Guidelines should refrain from providing speculative views on AI. Though it is a natural research goal to work on AGI, whether AGI is possible (and how long it will take to reach it) or not is an opinion that can be argued against or in favor of, but currently the answer is unknown.• Another long-term concern to be included is the monopolization of our attention based on AI technologies, with the unintended consequence of people getting addicted to some digital services. This may also result in the promotion of certain content based on AI-driven decisions which ends up in the reduction of quality of content being replaced with fake news and junk content. But this is more related to the business model of the service than with AI technology per se. Therefore, it is important to consider the business model in the trustworthiness assessment.

digital players; only they will be able to gather such an extensive and diverse amount of data, indispensable for the training of AI algorithms. No one will be able to compete as they will dominate the end-to-end ecosystem decide how AI evolves. This an ethical concern that should be raised within the assessment list in Requirements of Trustworthy AI: the requirement of Respect for Human Autonomy refers to protecting citizens in all their diversity from private abuses made possible by AI technology, ensuring a fair distribution of the benefits created by AI technologies. Certainly, in a monopolized data market where a few companies control the development of AI, and no alternatives are allowed, citizens could be subject to abuses from such companies. Unless abuses in the data gathering by the biggest digital players are avoided, allowing the emergence of alternative AI providers, benefits would not be fairly distributed among users and providers, neither geographies. At the end of this section the following paragraph could be added "Access to data and behavioral data is critical for the development and implementation of AI, and thus its monopolization in the hands of a very few companies could limit the emergence of AI solutions from other alternative players, thus enabling potential abuses on citizens and uneven distribution of AI benefits; situations critical data for the development of AI being monopolized should be avoided." • Architectures for Trustworthy AI do not seem to include the entire lifecycle of AI—including how data was collected and the business model of the whole system.• We recommend the addition of a reference to technical tools with the following capabilities:o Detection of correlations between sensitive variables and normal (apparently harmless) variableso Detection of bias in data sets (IBM, Accenture, Pymetrics, …)o Correction of bias: in the data set, and through the algorithm (using GANs, https://blog.godatadriven.com/fairness-in-ml) o Checking of the risk of re-identification of anonymized datao Visualization of the impact of false positives and false negatives on a certain domain

European citizens' personal data -a fundamental pillar for the development and improvement of AI- is controlled by non-EU businesses, having a framework imposing safeguards on AI that does only apply to EU based businesses will not benefit EU citizens nor EU business competitiveness. • If these Guidelines are becoming regulation to follow the "GDPR" model approach, it would be wise to first assess what is the impact of the application of GDPR on EU companies vs. non-EU companies and their respective competitiveness. This will provide critical learning for any possible future regulation on AI.2) Other general comments on the Guidelines:• Monopolization of data is a critical, immediate threat to the development and implementation of AI and could be an ethical question in and of itself. Since access to data, and more relevantly, behavioral data, is a requirement for the development of AI, and mostly all this data is going to be in the hands of a few companies, AI will be controlled by such companies. Indeed, the ability to track the behavior of billions of users worldwide, through the provision of multiple conglomerate services, is possible just and only for an extremely short list of global digital players; only they will be able to gather such an extensive and diverse amount of data, indispensable for the training of AI algorithms. No one will be able to compete as those few players will dominate the end-to-end ecosystem and decide how AI evolves. This an ethical concern that should be raised in the guidelines, for example within the assessment list in Requirements of Trustworthy AI in Section II: the requirement of Respect for Human Autonomy refers to protecting citizens in all their diversity from private abuses made possible by AI technology, ensuring a fair distribution of the benefits created by AI technologies. Certainly, in a monopolized data market where a few companies control the development of AI, and no alternatives are feasible, citizens could be subject to abuses from such companies. Unless abuses in the data gathering by the biggest digital players are avoided, allowing the emergence of alternative AI providers, benefits would not be fairly distributed among users and providers, neither geographies. At the end of this section the following paragraph could be added "Access to data and behavioral data is critical for the development and implementation of AI, and thus its monopolization in the hands of very few companies could limit the emergence of AI solutions from other alternative players, thus enabling potential abuses on citizens and an uneven distribution of AI benefits; situations where critical data for the development of AI is being monopolized, should be avoided." • It would be interesting to know what percentage of current AI applications complies and doesn't comply with the Guidelines, and give a few visible examples of each of those. This would also serve as a test case of how pragmatic and feasible the approach is. • Due to the relevance of data governance and the business model for trust in AI, we would suggest to highlight this also in the Executive Summary, and not only in the Rationale and Foresight section: "Trustworthy AI will be our north star, since human beings will only be able to confidently and fully reap the benefits of AI if they can

						trust the technology, the data governance and the business model.”• A minor terminological issue: instead of speaking about a “rule-based AI system”, it is better to call it “symbolic AI system” or “knowledge-based AI system”. Rule-based has a specific connotation in the AI world (if-then-else rules), but there are many other approaches (not falling under “learning-based”) that reason with explicit knowledge without rules.
Luc	Hendrickx	SMEunited	Given the complexity of the subject, its far reaching implications for all citizens and enterprises, the fact that it is more than giving an opinion, but that this needs reflexion, such a short deadline and launched in full X-mas periode (18 december-18 january, prolonged until 1 february) , this consultation cannot be considered as serious. It does also not respect the minimum standards for consultation. We will send however as soon as possible our position.			
Zoltán Kázmér	Szabó	MCOnet International	Egyetértünk az MI etikai iránymutatásokkal, de nem értünk egyet azzal, hogy Magyarországon nem magyar nyelven nyilatkozzon egy Magyarországon bejegyzett magyar cég.			

Johann Cas

Whereas I deeply appreciate the European commission’s initiative and the efforts of high level expert group to develop guidelines for trustworthy AI, I am also deeply concerned about some of the statements and tendencies in the draft version of these guidelines. The following comments only relate to these concerns, neglecting the many useful contributions and contents of the working document for stakeholder’s consultation. As my comments are brief and of a general nature, they are not split up according to the different sections of the working document, though some of them are related to specific chapters.The first critical comment is related to the selected title as such: there are only a few technologies imaginable that can be regarded as trustworthy on their own. In the case of AI the naming of “ethics for trustworthy AI” is in appropriate and misleading for several reasons. Depending on the concrete AI technology in mind, the results produced by these technologies are at least prone to statistical errors, some also show completely unexplainable (and unpredictable) behaviour. The labelling as guidelines for trustworthy AI contains at least implicitly the message that trust in these technologies is in principle justified as long as the developed guidelines are respected, neglecting the fact that it is the use of technology only, which could deserve this marking. In the case of AI this comment, which might be regarded as a linguistic sophistry, is doubly important. AI is threatening human autonomy and agency, as acknowledged in the working document; neglecting this fact in the very title is additionally endangering human agency. At least a renaming in the form of “ethics guidelines for trustworthy use of AI” or something similar should be considered. Otherwise the guidelines could become self-contradicting to one of the core principles mentioned in the document.A further critical comment is related to the condensing of the many fundamental rights and ethical principles touched by AI to 5 overarching principles. Such an attempt of simplification in general be useful to reduce complexity; in the specific case, it appears to be an oversimplification, containing the risk of resulting in non-operational principles. Do good or do not harm can hardly be contested, the elaboration of the concrete meaning of these requests is however by no means self-evident, nor the concrete measures to fulfil them. Whereas the origin of these combined principles (see footnote 5 of the working document) explains the composition of the five principles, the working document is missing these details. Listing the concrete composition of the combined principles would, however, not be helpful or sufficient. The blending of different fundamental rights and ethical principles rather conceal conflicts among them then helps to overcome them. These intentions to simplify and condense ethical principles is in addition in stark contrast to the very detailed requirements and assessment criteria listed in section B II and III.The attempt to provide all-encompassing requirements and assessment criteria is not criticised for missing important aspects in relation to AI. However, particularly relevant ethical issues for AI are mixed up with general requirements, like design for all. Requirements are included that I would not consider primarily as ethics issue, like

robustness, which could possibly be left to the market to be decided upon. The listed requirements have been relevant for any ICT innovation of the last decades; hence, they might rather conceal the critical issues pertinent to AI than to draw specific attention to them.The repeated use of the term reasoning for the results of AI is another potentially misguided attribution of features to AI. AI can rather be characterised by missing to provide reasoning, which makes human agency and attribution of accountability to human actors imperative for all AI applications having an impact on humans. There are no problems with using AI technologies to analyse the petabyte of data that are generated by high-energy experiments for instance at CERN. A central ethics issue is related to the use of rather stupid algorithms for decision-making on humans, for instance credit scoring or job applications, firming under the term AI. If AI is additionally attributed with reasoning, this will factually strengthen the position of AI in comparison to humans even more, regardless whether formal responsibility remains at human discretion.The provision of detailed assessment lists additional supports the impression that it is mainly a question of designing AI technologies, not the particular use that is made of these technologies, which constitutes the core of ethical issues. This impression is further reinforced by the invitation to provide thoughts on assessment lists for specific use cases, implicitly suggesting that the use cases are in principle ethically acceptable. For the fourth example, profiling and law enforcement, it is at least questionable whether the application of AI in this field is compatible with human dignity or democratic liberties at all. All the other use case examples might require ethical/societal considerations beforehand, e.g. how could AI contribute to more efficient and environmentally sound mobility or whether solidarity considerations should limit insurance premiums becoming adjusted to individual risks.Last but not least, the working document appears to overestimate the actual and potentially positive contributions of AI to solve the grand challenges our world is facing, missing to provide evidence for this positive overall evaluation. Whereas large and important positive potentials can be envisaged, past and current use of AI does not appear to support this judgement. Taking increasing economic inequality as an example, AI has rather contributed to it – e.g. in form of a key enabling technology of high-frequency trading on financial markets - but I’m not aware of making serious attempts to use AI to resolve imbalances on labour markets. On the political level, AI is rather threatening civil liberties and democratic systems than empowering citizens. On a global level, AI is rather supporting the establishment of worldwide monopolies than empowering consumers. Data based businesses possess unprecedented economic capacities, unparalleled political influence, powers to shape the results of democratic elections, unique possibilities to influence or to manipulate individuals in the information they receive or decisions they take. By disproportionately stressing the potential positive impacts and neglecting these already materialised threats, the working document in the current form might contribute to an inappropriate reliance on AI

							when tackling urgent problems of the EU and the world. By missing to mention these dangers and threats it also misses to analyse them and consequently to develop measures and policies to counter them. This leads back to the first critical comment: we are primarily not in need of a trustworthy technology but of making good use of opportunities offered by technology in the human interest and keeping human agency.Thank you much for considering my comments.
Sara	GHAZANFARI	ETNO - European Telecommunications Network Operators' Association	ETNO welcomes the draft "Ethics Guidelines for Trustworthy AI" launched by the European Commission's High-Level Expert Group (HLEG) on AI. We are delighted that the draft AI ethics guidelines place European citizens at the heart of AI development and use ("human-centric" AI), respecting fundamental rights, applicable regulation, and core principles that underpin the ethical purpose for AI. Several ETNO members have made public commitments to ethical principles governing the development and use of AI technologies. Our members have robust data governance programs, whose policies and procedures are also generally applicable to uses of data in AI applications and solutions. We support the guidelines' vision to create a culture of "Trustworthy AI made in Europe", which will not only protect and benefit individuals and the common good, but also enable Europe to become a globally leading innovator in AI, as it will generate user trust and facilitate AI's uptake. The establishment of a European approach to AI to foster competitiveness in the EU should be particularly emphasised. European values, enshrined in digital ethics, can represent a competitive advantage for the development of Trustworthy AI. Our understanding is that the guidelines are intended to help seize this potential. ETNO fully agrees with the acknowledgment that "no legal vacuum currently exists, as Europe already has regulation in place that applies to AI" and that the guidelines will not imply any form of regulatory intervention. The development, deployment and use of AI technologies are subject to a robust horizontal (and, in some areas like privacy, sector-specific) legislative framework that protects the fundamental rights and integrity of European citizens. Tightening the existing legal framework could stifle the European AI ecosystem rather than nurturing it, and ultimately let other regions of the word like China and the United States dictate the rules	ETNO supports a fundamental rights-based approach to AI ethics, underpinned by the families of EU's citizen rights described in the document. However, we have some remarks about the four identified principles that rest on fundamental rights (beneficence, non-maleficence, autonomy, and justice). Our main concerns are as follows: • The Principle of Beneficence: "Do Good" We encourage the HLEG to recognise commercial uses of AI technology as legitimate and beneficial. AI applications that increase efficiency and productivity have real positive impacts on society. A narrow application of this principle bears the risk of restricting companies' freedom to innovate. It could have undue adverse effects on the innovation capabilities of economic actors whose primary mission is not necessarily to improve collective wellbeing. It would also cause uncertainty with regard to existing applications that pursue legitimate business goals, but that do not clearly contemplate the "Do Good" principle. • The Principle of Non maleficence: "Do no Harm" Technology is a tool, not an end in itself. It is arguable whether a technology can be inherently "good" or "bad", or whether in principle all technologies can be regarded as ethically neutral and what determines their positive or harmful impact is their specific use. Therefore, any principles related to Good or Harm can only apply to the specific application and business model. Therefore, transparency regarding the application and business model of an AI system is more important than the transparency of that system's technological aspects. • The Principle of Autonomy: "Preserve Human Agency" ETNO supports the principle of autonomy, noting that it should recognise that different uses of AI call for different degrees and types of autonomy. This principle is largely reflected in the GDPR, whereby data	We think that this Chapter (and the guidelines more in general) should clarify where a distinction can be drawn between professional AI systems (i.e., used for businesses and public institutions) and consumer AI systems. The ethical frameworks and the measures to make a system Trustworthy may differ accordingly. There is a big difference between realising Trustworthy AI for a professional user (e.g., pilot, robotics operator, flight controller, etc.) and doing so for a regular person using an AI-based app for e.g., tax declaration or social security, though some applications could be less distinct, such as public sector use of AI in sentencing guidelines. ETNO would like to raise some comments regarding the identified requirements of Trustworthy AI. • Accountability: In our understanding, accountability goes far beyond redress and compensation for wrongdoings. Accountability is a much broader principle that requires an organisation to demonstrate respect of individuals' rights and compliance with applicable regulation and standards, as well as to be held responsible for its activities and their effects. Therefore, accountability mechanisms may include self-regulation instruments such as codes of conduct. • Data Governance: Data governance is a broader concept than what is reflected in this requirement. An organisation's policies, procedures, data protection officers, and training programs related to the use of data should all be relevant when assessing its approach to Trustworthy AI. Furthermore, we agree on the importance of datasets quality, but we are concerned that pruning biases away before engaging in training may in fact cause other, unintended biases to emerge. It may be preferable to identify the biases in the datasets before training, but to correct them ex post after the processing of the datasets has occurred. Particular attention should be given to the practice of	We do not have specific comments to this Chapter.	The draft guidelines mention that the HLEG will elaborate on four use cases in the final version of the document (Healthcare Diagnose and Treatment, Autonomous Driving/Moving, Insurance Premiums and Profiling and law enforcement). Even though telecommunication services are not contemplated in the list of use cases, ETNO would like to provide the HLEG with some elements describing the role of AI in our industry. We identify three main clusters of use cases enabled by AI: 1. Network Operations: As providers scale up their infrastructure by adopting network virtualization, software defined networks, cloud-based applications and 5G, AI becomes particularly crucial for efficiently operating the network. Network security and predictive maintenance of networks are just two of the most important use cases enabled by AI. 2. Customer Relationship: AI is key for enhancing CRM and customer experience. As much as many other sectors, telecoms are increasingly using customer service applications that rely on chat bots, virtual assistants, and personalized content and offerings in real time. 3. New Products: AI systems are important for the development of new, data-driven services and products. Several telcos are investing in the creation of "platform ecosystems" for their clients, largely powered by AI. An example is data management platforms offered by telcos, where their customers can store, share and use data in a secure and privacy-protective manner. On a separate note, we would like to comment on the definition of AI that is provided in the addendum to the guidelines. This definition does not seem fully accurate and would deserve further consideration. For instance, the goal an AI system is tasked with meeting may not be necessarily complex and an AI system is not necessarily

of the game.
Nevertheless, we recognise that some elements of the existing framework (e.g. cybersecurity) may need adjusting to the new challenges brought by AI. In this respect, ETNO agrees with the statement that "different situations raise different challenges". Different AI-based systems may have a different impact on the rights of individuals at any stage of their life cycle. Therefore, we recommend embedding a clear "risk-based approach" in the guidelines and any possible future initiatives on AI, recognising that the requirements and methods for achieving Trustworthy AI should vary depending on the specific AI system's application.
We also encourage the HLEG to ensure that the document does not contradict EU law by introducing novel terminology or by reinterpreting specific, well-established legal concepts and obligations especially related to the General Data Protection Regulation (GDPR). Furthermore, we also suggest that the guidelines clarify what terms like "wellbeing and the common good" mean according to the EU understanding based on fundamental rights.
Finally, although it is clear that the guidelines will be voluntary and non-binding, it is less clear what the practical implication of their formal endorsement by stakeholders will be. Most notably, it is unclear whether any benefits or duties will be attached to the formal adoption of the guidelines, and how stakeholders' compliance with them will be scrutinized. It is also unclear how endorsing the guidelines will affect existing self-regulatory initiatives, such as guidelines and codes of conduct, already implemented by individual organizations. We ask the HLEG to elaborate in further detail on the concrete functioning and effects of the future mechanism for endorsement. This is crucial for ETNO, as many European telecom operators have already launched their own guidelines, manifestos, dedicated work streams or committees.

subjects have the right not to be subject to a decision based solely on automated processing (Art. 22) and have a right to object to most forms of processing of their data at any time (Art. 21). It is then important that the principle of autonomy as described in the guidelines be consistent with the existing legal framework. For instance, footnote 13 could be interpreted as an extensive right to object to any AI-based data processing in the working environment, beyond the letter of Art. 88 GDPR on processing in the context of employment. The flexibility and balancing of interests inherent in the GDPR are a valuable reference in this context.
• The Principle of Justice: "Be Fair"
Besides the concept of fairness and the importance of redress mechanisms and remedies (which are already provided for by the GDPR), we support the concept that human agents are ultimately responsible for AI-based decisions and their impacts on individual rights. Identifying the person(s) and/or role(s) responsible for a given system should be part of every developer and implementer "accountability" mechanisms.
• The Principle of Explicability: "Operate Transparently"
We agree that AI systems should be as transparent as possible for users, to provide them an understanding of how decisions affecting them are taken. However, we recommend applying the proportionality and risk-based approach principles to explicability, whereby the degree of insights required would depend on the complexity of the system as well as on its impact on individuals' rights.
Furthermore, we would like to comment on the statements that "informed consent is a value needed to operationalise the principle of autonomy in practice" and that "in order to ensure that the principle of explicability and non-maleficence are achieved the requirement of informed consent should be sought". We would like to remind that, according to the GDPR, user consent is just one of six legal bases for processing personal data. With regards to automated decision-making, the data subject has a right to object when the processing produces legal effects or significantly affects him or her. Therefore, we recommend not considering consent as the panacea to ensure the respect of the ethical principles at hand. Depending on the context, other legal basis may be equally or more suitable for ensuring transparency, explicability, non-maleficence and accountability of an AI system.
Identification without consent per se is not unethical and does not automatically imply a threat for individuals.
Moreover, regarding "the usage of "anonymous" personal data that can be re-personalized", the potential for re-identification depends on the technical means used to anonymise or pseudonymise personal data as well as the way data is clustered, packaged and processed thereafter.
As to "Covert AI systems", we are not convinced that these systems represent a critical concern as such. That will depend upon the function the system provides and the context in which it operates. Appropriate transparency measures towards users of AI systems are key; nonetheless it should be considered that, in some context, individuals that know they are interacting with a

data labelling. We also have doubts about the description of how anonymisation should not hamper a proper division of datasets for training and test. Anonymisation is not per se linked to which data is used in training and test, as long as the same data is not used in both sets; two different pictures can easily be split so that one ends up in training and the other in testing, which has nothing to do with the process of anonymization. Finally, this section could elaborate on the legal grounds available for processing personal data.
• Design for all: We agree that AI systems should in principle be accessible by all citizens. We would also note that some AI-based products and services may target one or some specific groups (e.g., age-specific or gender-specific) while not barring everyone else from technically accessing that system. "Positive discrimination" is not automatically in contradiction with this requirement; for instance, an AI-based product may be specifically designed for disabled people.
• Governance of AI Autonomy (Human oversight): We welcome the risk-based approach attached to this principle. We believe that a clear designation and communication of the person(s) and/or role(s) responsible for a given system should be a key part of good governance.
• Non-discrimination: As already mentioned, positive discrimination is not necessarily unethical and may even be necessary to reach an objective. For example, medical researchers may need to study a component of the population that have specific characteristics, and use AI to extrapolate this sample by excluding the rest of the population.
• Respect for (& Enhancement of) Human Autonomy: It may be difficult for an AI system to protect citizens from abuses by design. Systems should include processes to avoid their misuse, but it is very hard to prevent any governmental or business abuses that depend on the actual usage of the technology.
• Respect for Privacy: We suggest that the guidelines highlight the importance of effective technical and organisational measures that mitigate the privacy risks for individuals, such as the "pseudonymisation" of personal data.
• Robustness: Security and resilience to attacks are fundamental prerequisite of robust AI systems. We suggest that the guidelines expand on what mechanism could be implemented to ensure high cybersecurity standards for AI systems (e.g., "security-by-design"). We recommend assessing the relevance of the regulatory framework for operators of critical infrastructure (i.e., Directive on security of network and information systems) for AI systems.
• Transparency: We reiterate that explainability should be guided by the principle of proportionality and the risk-based approach.
With regard to the technical and non-technical methods to achieve Trustworthy AI, we have the followings remarks about the technical methods described by the guidelines:
• Architectures for Trustworthy AI: Trustworthy AI should not only be ensured by "formulating rules, which control the behaviour of an intelligent agent, or as behaviour boundaries that must not be trespassed", but also through mechanisms

designed by a human, but by another machine.
We recommend that a revised definition of AI features the following criteria:
• exclude software systems based on traditional and determined algorithms that are clearly not based on AI;
• take into account that the AI algorithm takes decisions as a consequence of the application of advanced analytical techniques (i.e., machine learning and deep learning) to solve problems;
• require strict ethical scrutiny of an AI system only when its purpose may constitute a risk to individuals' fundamental rights (risk-based approach).

			machine will behave in a different way that hinder the objectives of the system (e.g. in medical research). Finally, we believe that providing examples of potential longer-term concerns at this stage could be premature and could fuel unfounded worries. For instance, Artificial Moral Agents (AMAs) should not per se pose a threat as long as these have been trained within a given and acceptable ethical framework; on the contrary, AMAs might well be considered one of the few technology principles for developing ethical AI in practice. Additionally, whether self-improving Artificial General Intelligence (AGI) is possible is still a matter of speculation. These are subjects that could be considered in an eventual follow-up phase of discussions.	enabling operators to deactivate and stop AI systems at any time. • Traceability & Auditability: The meaning and the objectives of traceability and auditability for the purpose of these guidelines should be clarified, bearing in mind the context and the application (professional vs. consumer) of an AI system. Producers and developers of AI should keep track of the decisions made and the information fed to the system also in order to enhance the quality of decisions. • Codes of Conduct: The headline is misleading, as there is more to ensuring an organisation's adherence to ethical principles than just codes of conduct. We suggest renaming the section "Corporate Governance". Additionally, we would like to suggest further technical methods to achieve Trustworthy AI: • Responsibility: As already mentioned, AI systems should have a responsible person or role that takes decisions regarding the system and monitors its operations. Responsibility should be present at every stage of the system's lifecycle. • Pseudonymisation: Pseudonymisation of personal data enables data processing in a privacy-friendly manner but, contrary to full anonymisation, it preserves the necessary identifiers that allow to repeatedly merge large amounts of data from various sources over time while eliminating the direct link between data and data subject. The EU has embraced pseudonymisation as a privacy-friendly technique in the GDPR.		
Matteo	Quattrocchi	BSA - The Software Alliance	BSA | The Software Alliance welcomes the opportunity to offer thoughts on the High Level Expert Group (HLEG) Draft Guidelines. BSA is the leading advocate for the global software industry before governments and in the international marketplace. Our members are at the forefront of software-enabled innovation that is fueling global economic growth, including cloud computing and AI products and services. BSA members include many of the world's leading suppliers of software, hardware, and online services to organizations of all sizes and across all industries and sectors. BSA members have made significant investments in developing innovative AI solutions for use across a range of applications. As leaders in AI development, BSA members have unique insights into both the tremendous potential that AI holds to address a variety of social challenges and the governmental policies that can best support the responsible use of AI and ensure continued innovation.The formation of the HLEG is a unique opportunity for Europe's leading experts from industry, academia, and civil society to help the European Commission develop a "coordinated approach to make the most of the opportunities offered by AI and to address the new challenges that it brings." We agree with the Commission that the success of such a framework will turn in large part on whether it fosters an "environment of trust and accountability around the development and use of AI."The HLEG requested comments on the potential concerns raised by AI. BSA agrees that it is important to consider potential short and long term concerns. BSA supports efforts to deploy AI responsibly. We urge the HLEG to	BSA supports discussions of high-level ethical principles pertaining to Artificial Intelligence development, especially within the frameworks of the EU Treaties and Charter of Fundamental Rights. While such an exercise is important to begin a principles-oriented debate at the European level, it is also important to underline that AI development and use is happening globally. The EU would greatly benefit from considering the existing best practices, developed by public and private sector, at the global level. It is important both that EU values and priorities are fully taken into account at the international level, and that international standards and principles are incorporated in any European effort. While the Guidelines will need to reflect European values, we must also keep in mind that AI will be developed and deployed in an international context. If European good practice or European guidelines are too draconian, prescriptive or overly rigid, AI will be developed elsewhere and other geographies will reap the benefits of AI innovation while Europe is left behind. The international standards community is beginning to address many of the issues raised in the Guidelines. BSA recommends that European authorities and industry fully engage in these international efforts.As a global organization, BSA has also developed a set of 5 principles for building confidence and trust in AI systems, which are consistent with the 5 principles the HLEG suggests in the Draft Guidelines. The BSA principles are:a) Fairness—considering measures to evaluate AI systems to help recognize improper or unconscious bias; b) Accuracy—acknowledging the importance of data	Trustworthy AI is a critical objective and BSA supports the HLEG efforts to provide a general framework for practitioners to achieve it. As the HLEG has rightly noted the "Guidelines are not meant to stifle AI innovation in Europe, but instead aim to use ethics as inspiration to develop a unique brand of AI". We therefore recommend a light-touch approach on many aspects of the Guidelines, in the interest of working with practitioners around the world to develop a framework that supports innovation while providing workable guiding principles. To that purpose, BSA suggests streamlining the below list of 10 requirements, and aggregating some of the requirements to ensure that the list is of easier use.In particular, we would like to provide the following comments to the 10 requirements the HLEG identified in the Guidelines:1) Accountability: BSA agrees accountability in development is needed. We caution that this is among the areas in which the mechanism has to be sufficiently flexible to accommodate different use cases and means of deployment. 2) Data governance: BSA believes that a trustworthy AI should respect the principles of Accuracy (i.e. acknowledging data quality and, where feasibly, identifying sources of error in data inputs and system outputs), Data Provenance (i.e. considering measures that could facilitate evaluation and documentation of data used to train AI systems, how those data are collected, and how data is used over time within AI systems, consistent with any other data retention obligations) and Fairness (i.e. considering measures to evaluate AI systems to help recognize improper or unconscious	BSA strongly supports any effort by the HLEG, and subsequently by the Commission, to involve a broad audience of practitioners in developing best practices and guidance for assessment instruments. As the HLEG rightly notes in the Draft Guidelines, any assessment instrument, tool or best practice will have to be considered in context, and given the specific purpose of each AI application. We also support the HLEG's efforts to provide such guidance, which is important to enable innovators to understand in a practical way how ethical principles can be deployed.As a global organization whose members are at the forefront of AI development, BSA has developed a list of practices for responsible AI deployment. These practices, provided below, are necessarily high-level because of the numerous use cases and deployment models: • Conducting in-house testing and evaluation of AI systems to ensure they meet their specified goals; • Developing guidelines and providing necessary resources to developers to help evaluate fairness and guard against improper bias; • Identifying persons with relevant expertise who are responsible for addressing significant problems identified with operating AI systems; • Ensuring subject matter experts, especially those with knowledge of the policy landscape in which the AI system will be deployed, are available to assist computational scientists in the design and implementation phases; • Providing descriptions of procedures used to assess the quality of data inputs and address errors identified in outputs; • Providing general descriptions, where appropriate, of training datasets that AI systems use to learn; •	For more information about BSA's work on Responsible Artificial Intelligence, please visit ai.bsa.org

be cautious in recommending policy action now based on speculative concerns, however. The Draft Guidelines make the important point that "it should be noted that no legal vacuum currently exists, as Europe already has regulation in place that applies to AI". BSA believes that building trust in Artificial Intelligence system is one of five key pillars that we have identified for the development of Responsible AI. These pillars are:1) Building Confidence and Trust in AI Systems: please refer to the list below (Chapter I) of five principles for more information on this pillar.2) Sound Data Innovation Policy: The exponential increase in data has fueled advances in machine learning and AI. Facilitating the development of AI requires sound data innovation policies.3) Cybersecurity and Privacy Protection: BSA advocates for policies that strengthen enhanced security measures and respect informed consumer choices while ensuring the ability to deliver valuable tailored products and services.4) Research and Development: Investment in education, research, and technological development will be integral to continued development of AI technologies and global economic growth.5) Workforce Development: The increasing use of and demand for technology is creating new types of jobs, in every sector of the economy, that require an evolving set of skills.The discussion fostered by the HLEG around Trustworthy AI is a fundamental step in ensuring that the EU has a leadership role in innovative technologies. At the same time, excessively prescriptive guidelines can be counterproductive in a field that is in rapid development, such as AI. The HLEG's work is an important step in initiating a larger debate across the EU on Artificial Intelligence, and BSA supports its ethical framing. BSA recommends establishing a clear process and timeline to update the guidelines with stakeholder involvement to ensure the HLEG recommendations stay relevant and in line with technological development. While the AI Alliance is certainly an effective platform for engagement, BSA recommends the creation of additional fora at the EU level to ensure stakeholders involvement, as well as establishing a timeline for a short-term evaluation of the guidelines. BSA welcomes the Draft Guidelines as an excellent effort to strengthen the EU's role as a global leader in ethical and responsible development of innovative technologies. Overall, the Draft Guidelines seek to strike the right balance between complex ethical dilemmas and the need to support AI development in the EU. In their current formulation, they are not merely a compilation of values and principles, rather they also provide guidance on how to implement these principles. BSA considers the Draft Guidelines an excellent step and foundation for the AI conversation at the EU and global level. Moreover, the HLEG's recognition that Trustworthy AI has two components – ethical purpose and technical robustness - is an important guiding principle for any work done on AI around the world.At the same time, BSA also recommends that the HLEG takes into consideration similar international efforts in the space of AI and ethics, to ensure that the EU can contribute to the global discussion and development of AI tools. An international perspective on how to realize trustworthy AI, and assess it, would ensure quality and, where feasible, identifying sources of error in data inputs and system outputs; c) Data Provenance—considering measures that could facilitate evaluation and documentation of data used to train AI systems, how those data are collected, and how data is used over time within AI systems, consistent with any other data retention obligations; d) Explainability—exploring how to provide reasonable explanations of how AI systems operate; and e) Responsibility—considering whether processes are available to address unexpected issues that may arise after AI products and services are deployed.At a high level, both the Ethical Principles discussed in the Draft Guidelines and the BSA principles stress the importance of designing and using AI as an understandable tool to aid human decision making and improve economy and society.BSA welcomes the HLEG recognition of the diversity of AI applications, and the importance in avoiding the creation of a "one-size-fits-all" regime. We believe that contextual considerations merit greater attention in the Guidelines. The degree of risk of individual or societal harm, and the potential severity of such harm, will vary enormously depending on the specific AI application at issue. In fact, many of the ethical issues identified in the Guidelines only arise for AI systems that have a consequential – or meaningful – impact on individuals. BSA therefore urges the HLEG to make clear at the outset of the Guidelines that the recommendations should be tailored to each specific implementation of AI depending on a careful and thorough risk assessment.Engagement at the principles-level is an important step in strengthening trust in AI tools. To that end, BSA commends the HLEG's efforts to ensure that the Draft Guidelines take into account the multitude of diverse applications of AI, as well as the technical considerations related to enacting ethical principles in this space. BSA encourages the HLEG to recognize more explicitly that AI policy involves trade-offs, and therefore a risk-based approach, tailored to the circumstances, will be necessary. BSA also recommends that the Guidelines adopt a broad understanding of beneficence. In fact, AI can be a tool to improve wellbeing, but it can also serve more neutral objectives whose direct individual or social benefits are less clear.

bias). The Guidelines' concept of "data governance" should be broader and reflect the fact that governance structures necessary to develop AI ethically include a broad range of engineering and design practices as well (e.g. access controls, systems documentation), BSA therefore urges the HLEG to recognize that data governance is complex in practice and will need to be tailored to individual scenarios.3) Design for all: BSA supports broad access to AI products, in particular as many applications of AI will greatly benefit underserved portions of the population. To that end, we would suggest including recommendations for public sector support to industry to develop products with high accessibility. Nevertheless, we would also caution against overly prescriptive requirements in the development and design phase, as flexibility in innovation is an integral part of any development process.4) Governance of AI autonomy (Human oversight): BSA agrees that human oversight is an important principle, given the diversity of AI tools, and the different technical considerations they would entail (e.g. in machine-to-machine applications), BSA would suggest to also consider context and purpose of an AI technology with this requirement.5) Non-discrimination: limiting bias and unfair discrimination are fundamental objectives. BSA believes that trustworthy AI should respect the principle of Fairness (i.e. considering measures to evaluate AI systems to help recognize improper or unconscious bias) and has as well put forward a number of best practices recommendations to limit the effects of unfair bias in AI development (please refer to our comments on Chapter III). At the same time, BSA would like to stress that measures in place to limit bias should not be considered absolutely foolproof. In some instances, it may be necessary and/or intended to consider certain individual characteristics (e.g. in the healthcare sector, diagnosis tools might need to consider age, sex or personal background as factors for diagnosis, as they might lead to higher propensity for some diseases or different reactions to cures). BSA would stress the importance of highlighting the need to protect against unfair discrimination and bias.6) Respect for (& Enhancement of) Human Autonomy: BSA agrees that fundamental and constitutional rights need to be safeguarded with the progressive deployment of AI technologies. While the principles defined in Chapter I are designed to improve and guide AI development, it is important to stress that AI will function in an already strong rule of law system within the EU, and will be designed to respect and strengthen that system.7) Respect for privacy: BSA fully agrees that trust and privacy are foundational to the development and adoption of AI. Beyond the necessary full compliance with GDPR, BSA promotes best practices globally that increase the transparency of personal data collection and use; enable and respect informed choices by providing governance over that collection and use; provide consumers with control over their personal data; provide robust security; and promote the use of data for legitimate business purposes. 8) Robustness: BSA is a strong advocate of data accuracy, resilience and cybersecurity. The more complex a system, the more

Developing mechanisms for consumers to request information, obtain guidance and address potential concerns; • Continuing monitoring after product release to detect and address unintended outcomes; • Providing visual aids and/or plain language explanations that communicate important facts about AI systems and their operation; and/or • Supporting continued research and analysis of transparent modeling. The HLEG asks specifically about how an assessment would work in four "use cases": (1) healthcare diagnose and treatment; (2) autonomous driving/moving; (3) insurance premiums; and (4) profiling and law enforcement. BSA agrees with the draft Guidelines that precise questions relevant to assessment of any particular AI system will vary depending on the use case. The difficulty is that these use cases are themselves broad categories. There are numerous different uses and deployment models for AI within each category, that have different levels of risk based on the nature of data sets, the time for human intervention, and numerous other factors. For these reasons, BSA developed the more general list of best practices above, and urges the HLEG to consider a similar, flexible approach.

that the EU remains competitive on the global markets, whilst contributing to strengthen trust in new technologies.On a more general point, BSA recommends a more positive approach to Artificial Intelligence. The HLEG correctly noted the tremendous potential AI has to spur economic growth across every industry sector, improve human decision-making in ways that will make the world more inclusive, and enable cutting-edge breakthroughs on vexing social challenges such as climate change and cancer research. BSA therefore recommends:- Creating additional instruments for meaningful and routine stakeholder consultations after the publication of the final guidelines;- Strengthening the role of the members of the AI Alliance platform and create additional means of stakeholder involvement;- Establishing short-term and long-term timelines for evaluation of the guidelines;- Establishing a clear process to amend and update the guidelines to ensure they remain relevant and in line with technological development;- Ensuring that the guidelines are informed by and contribute to international efforts.Finally, we note the new definition of AI provided by the HLEG. BSA appreciates the Guidelines' thoughtfulness in proposing a possible definition of AI. We note, however, that many solutions in use today that are described as having an AI component make connections, reveal correlations, or provide other insights that humans then use to decide on a course of action, but do not necessarily decide "the best action(s) to take" as stressed in the Guidelines' definition. BSA is still considering the implications of the proposed definition and may provide further input on it after conducting a more in-depth analysis.

important these principles and practices become. As mentioned above, BSA has developed a set of principles and best practices to ensure trust in AI tools, and in particular strongly believes that Accuracy, Data Provenance, and Responsibility should be the guiding principles for robust and trustworthy AI systems. With regards to cybersecurity, BSA has developed a wealth of materials to promote cybersecurity awareness, while protecting privacy and safety (for more on this, please refer to bsacybersecurity.bsa.org).9) Safety: BSA fully supports safe systems and believes that trust can only be earned through safety in practice. Safety is a fundamental component of trust in AI tools, and BSA is fully committed to the highest standards in AI development and deployment. 10) Transparency: Explainability is a key principle to ensure trustworthy AI. In particular, BSA believes that explainability (which is a more accurate term to use in providing an explanation of the AI system's approach in useful terms for the user) will inevitably vary due to context and purpose, it is also important to develop strong best practices and support information for all users of AI tools. It is as well fundamental to acknowledge that any effort in the field of transparency will need to take into account the developers' ability to innovate and provide cutting-edge services. BSA believes that these efforts would be better developed if led by industry and developers as AI tools are deployed. In addition, achieving transparency can be complex and highly dependent on a host of variables, precluding a "one-size-fits-all" approach. When it comes to technical methods to achieve Trustworthy AI, and particularly with regards to traceability and auditability, BSA believes that the nature of auditability will be heavily context dependent. In complex scenarios, third party auditors and expert controls will be more effective for technical support. In still other scenarios, internal organizational auditing and controls may suffice. In light of this, the Guidelines should do more to acknowledge that effective auditing, depending on the context, can include any of those mechanics.BSA is a strong supporter of efforts to create risk-based regimes that support solutions to significant technical and operational issues in new technologies. The HLEG should consider advancing the use of risk assessments for AI as a tool for companies to interpret how technological, operational, and policy controls, requirements and standards can support implementation of Trustworthy AI. Guidance should also embody a risk-based approach to deciding where companies should most effectively focus their efforts. Risk management is at the core of advancing trustworthy AI. Those developing or implementing AI should be responsible for conducting appropriate, robust risk assessments. Identified risks should be mitigated through effective safeguards, such that the benefits of the AI implementation outweigh the residual risks. This is fundamental to ensuring that users and other stakeholders are protected and safeguarded.

Anonymous	Anonymous	Anonymous	I give this feedback with utter respect to the people who have invested their time and energy into the work to date. My comments are not designed to be derogatory to the authors but to hopefully maintain or raise standards. ----------"The AI HLEG is convinced that AI holds the promise to increase human wellbeing and the common good but to do this it needs to be human-centric and respectful of fundamental rights."Comment:While I'm an optimist I feel that this, particularly as an opening statement and main positioning statement, is naive. If I were a Hollywood writer I would use this type of setup statement in order to make the humans look silly when the fall happens.----------"We therefore set Trustworthy AI as our north star"Comment:I humbly say that this statement shows a fundamental misunderstanding of what AI is and will evolve into. You cannot by definition create an intelligence and expect it to obey a subjective human trait like 'Trustworthy' The definition of Intelligence (Google) is "The ability to acquire and apply knowledge and skills." It should be understood that we are creating this ability with AI. Making a law that compels business and individuals to make good trustworthy and ethical decisions when building AI is only a patch on a leaky dam. As pointed out creating an intelligence gives it the ability itself to be trustworthy or not.----------(I) Ethical Purpose. This Chapter focuses on the core values and principles that all those dealing with AI should comply with. These are based on international human rights law, which at EU level is enshrined in the values and rights prescribed in the EU Treaties and in the Charter of Fundamental Rights of the European Union. Comment:AI itself as an intelligence independent of it's creator will need to subject to EU law					On Darwinism, The alpha species and being 'top of the food chain'"AI technology" is not like existing traditional industrial age technology. Today's most advanced software is just an evolved loom from the industrial revolution. It it a tool which humans use for efficiency. I describe it this way for context.I propose disassociating AI with the term "technology" in order to give it a mental model separation from our simple tools. AI is (as named), an Intelligence. We as the Alpha species on this planet are inviting a new Intelligence to the top table and as leaders we have a responsibility to the entire human race to make wise decisions and not be naive in the face of the excitement that this phenomenon of evolution is creating.
Henrik Palmer	Olsen	University of Copenhagen, Faculty of Law						While there is much to agree with in these proposed guidelines, we have identified what we consider to be a number of generalizations and assumptions that we do not find persuasive and which, in our view, should be addressed on redrafting. For instance, one of these generalizations is found very early on in the document, where it is claimed, “on the whole, AI’s benefits outweigh its risks”. Given the technology’s lack of real world scenario testing, this is a large and mainly undocumented assumption. Further into the document, we find that they at several places emphasize that “Trustworthy AI” has two components: 1) ethical purpose, and; 2) technological robustness While we fully agree with the need for AI solutions to be both ethically and technologically sound, we take issue with the way the document is phrased in regard to an ‘ethical purpose’. On p. 2, it notes that Trustworthy AI, in its development, deployment, and use should “… respect fundamental rights and applicable regulation, as well as core principles and values, ensuring an “ethical purpose”. While we agree that AI must be implemented in society in a way that respects basic ethical principles, we think that it may be inappropriate to phrase this as a requirement that AI should have an “ethical-first” focus. While we do not disagree with

the need for ethical reasoning applied to the field of AI technology where this technology connects with human agency, it is our firm belief that the achievement of ethical defensibility is best realized through legal procedure.
AI is to a large extent demanded by the private and public sector, either because it enhances efficiency or because it allows for companies and/or public institutions to enhance their knowledge and thereby act in smarter ways. The development of AI is largely driven by the wish to meet the request for technological solutions that enhance efficiency or a knowledge base. Therefore, we think it is a little misleading to focus on "ethical purpose" as the driver of AI development, which in reality is as helpful as it is vague. The market demands are, and should continue to be, the drivers of this development, but the markets should not be allowed to operate freely. Instead, regulation of AI should be made to ensure that the technology is used only for purposes that are ethically sound and oversight by public agencies should be put in place to ascertain that such regulation is respected to its full extent.
In light of this, we emphasize that specific legal regulation rather than "ethical reasoning" (see p. 2) should be at the center of ensuring the ethically defensible use of AI. It is incumbent on the Commission to draft guidelines that highlight the regulatory framework. Ideally, it should go far beyond a mere regurgitation of GDPR principles, which see citizens merely as data subjects (which is just one aspect of AI regulation), and go beyond the model of mere consent (see p 10).
The document, to some extent, recognizes the need for a legal framework for AI technology, in that it emphasizes that AI in its development, deployment and use should respect fundamental rights and applicable regulation (p. 2). However, a legal framework for AI entails more than respect for fundamental rights and existing applicable regulation. A legal framework for AI that is to meet both the efficiency and knowledge gains that drives the demand for AI and simultaneously meets the ethical requirements considered by the High Level Expert Group requires a more detailed commitment to legal analysis of AI as used in various domain and institution specific settings. Without a robust description of the specific gaps in legislation and what is needed to close them, this document pays no more than lip service to legal regulations' role in protecting citizens. This, we think, is a big problem, since free-floating ethical reasoning will have no bearing on the ability to effectively formulate, enact and enforce principles and rules pertaining to the proper use of AI in society.
Lastly, in regard to the standards applied to AI ethics and AI regulation, we would like to point out that the technological quality and ethical soundness of AI should not be judged against some kind of ideal or perfectionist ethical standard. AI is a technology that is being introduced to make human labor more efficient or better. The best we can demand from AI is that it is successful in this regard while not undermining the quality of human lives or the legal commitments made to ensure that quality.

Prof. Henrik Palmer Olsen, Dr. Jacob

Matthew	Newman		Chapter I - Section 4 – Introduction – Paragraph 3Suggest remove recommendation for ethical expert here. This may be viewed as a sales pitch, and is the only implementation recommendation in this whole chapter.Chapter I - Section 4 – Subsection 'The Principle of Non maleficence: "Do no Harm"' – Paragraph 3Suggest a broader handling with environment as an example. The key question here is the trade-off between short term benefit causing long term harm. Recommend we have the ambition to recognise the rights of future generations decided by current activity, requiring we don't only consider current stakeholders but future stakeholders too.Chapter I - Section 4 – Subsection 'The Principle of Autonomy: "Preserve Human Agency"' – Paragraph 1"Autonomy of human beings in the context of AI development means freedom from subordination to, or coercion by, AI systems" – this appears to infer attribution of intent to the system. Suggest language to refer to the legal party controlling the AI rather than inferring the system is taking decisions without the control of a sentient party. Chapter I - Section 4 – Subsection 'The Principle of Autonomy: "Preserve Human Agency"' – Paragraphs 1 & 2"human agency - a right to opt out and a right of withdrawal. Self-determination in many instances requires assistance from government or non-governmental organizations to ensure that individuals or minorities are afforded similar opportunities as the status quo." – recommend the requirement that the right to opt-out does not lead to a any penalty to human agency – comparable example: internet banking marginalising elderly people's access to financial independence. Opting out should not be to the detriment of one's rights.Chapter 1 Section 5.1 Thoughts: Images that can be used to link personal data (either volunteered or not) to other data that has not been volunteered should be treated as the equivalent of the total data set made from the linking. Example: a data set with identifiable personal data linked to anonymous records hospital visits should be treated as personal medical data.Minimisation principle should be applied in crafting consent: the individual cannot be asked to consent to non-specific uses of their	Chapter 2 Section 1 Subsection 2. Data Governance "The datasets gathered inevitably contain biases, and one has to be able to prune these away before engaging in training. This may also be done in the training itself by requiring a symmetric behaviour over known issues in the training set." – recommend to provide guidance on discriminating bias. Not all bias is bad, stripping all bias is a fools-errand and actually will likely defeat the purpose of the AI. The question is how a developer could distinguish bias affecting the rights and principles.Chapter 2 Section 1 Subsection 4. Governance of AI Autonomy (Human oversight)Suggest mention that the oversight needs to be from an individual empowered, able and who can realistically recognise and act upon deviation signals (i.e. avoid human moral crumple zones)Chapter 2 Section 1 Subsection 6. Respect for (& Enhancement of) Human Autonomy"Systems that are tasked to help the user, must provide explicit support to the user to promote her/his own preferences, and set the limits for system intervention, ensuring that the overall wellbeing of the user as explicitly defined by the user her/himself is central to system functionality." – I have concerns that this wording encourages echo-chambers; suggest we include the concept that preference is promoted but context/alternative is always provided. Chapter 2 Section 2 Subsection 2I have more extensive feedback on this section that I am happy to contribute through a better channel than a text box. Two keys point:Accountability Governance – suggest a bit more rigour (e.g. inclusion in the policy framework of an organisation). Having a CXO for data or ethics brings less value if they are not bound to policies that the board are judged on. This is the route to internal controls frameworks.Codes of Conduct – suggest you add a recommendation for a change program of work to embed these, the inclusion of such standards as part of HR review process, training, onboarding, etc	I have general feedback on this section, so will omit comments on individual section or sub-points. As it's currently constructed, I feel it might not be the most useful tool for assessment, though the contents are very positive. It's challenging to understand how a company could put this into practice. I would recommend breaking the list into steps concerning policy and governance; practice and standards; product/service development; system development; etc. This would make it a more useful tool. I'm happy to be involved in the discussions on practical implementation steps, recognising that most change is more successful when expressed in the terms of the affected party rather than that of those defining the change.	Suggest checking language to explicitly recognise that functionality, and therefore risk of deviation from ethical goals, alters in AI systems during usage through learning. I feel that much of the focus in the document is on developers rather than those roles who will have a better understanding of the market, potential users, etc. Suggest that more focus should be given to distributed development. Much of the document assumes a traditional enterprise with line management hierarchies and governance. Not covered are distributed development, app store development, open source, white label, etc.

							data, but only to specific processing with explicit mention of the outcome and potential harm (compare sheet included with pharmacy medicines).Chapter 1 Section 5.2I would suggest what is important is not knowing whether it's an AI or not, but when the entity with which we are interacting is representing another party other than itself, and whether the views, intent, etc expressed are not its own. This will naturally include androids, etc. Chapter 1 Section 5.3 Requires opt out with no detriment. The use of AI in scoring calls to question whether we would like to preserve a more fundamental human trait. Should we accept that humans have varying levels of ability to function in society and that it is part of being human to provide effort to compensate for others in an attempt to provide equality of outcome? Examples: how much insight should insurance have? Should credit card companies know credit history of family? Chapter 1 Section 5.5 Add to this group an increase in edge AI utilising partially trained networks (where the user completes the "training" e.g. phone based assistant), distributed AI (basement developers), AI DAOs. I struggle to do this topic justice in a single text box. Happy to discuss further.		
STEPHANIE	ALVES	CCBE - Council of Bars and Law Societies	The Council of Bars and Law Societies of Europe (CCBE) represents the bars and law societies of 32 member countries and 13 further associate and observer countries, and through them more than 1 million European lawyers. The CCBE appreciates the opportunity to take part of the ongoing "Stakeholders' Consultation on Draft AI Ethics Guidelines", being a member of the European AI Alliance, and has carefully taken into consideration the draft paper as a result of the discussion currently taking place between the 52 experts of the High Level Expert Group (HELG). The issues and principles set out in this paper are all significant aspects to consider around the use and development of Artificial Intelligence (AI) systems. Due to the tight time schedule of the consultation, the comments submitted below are general remarks for the purpose of presenting a preliminary analysis of the issues set out in the Draft Ethics Guidelines for Trustworthy AI and are therefore still subject to any position being developed at a later stage by the CCBE on this topic.• We understand that the HLEG paper (hereafter 'this paper') is a starting point for the discussion on a "Trustworthy AI made in Europe". As a general remark, we appreciate that the experts of the HLEG have provided a comprehensive view on the way how to achieve a "Trustworthy AI" by defining an ethical framework for achieving it. • The authors not only propose a wish list, but also methods for effectively achieving the established goals. The list of methods includes both technical characteristics and	Chapter I deals with ensuring AI's ethical purpose, by setting out the fundamental rights, principles and values that it should comply with. • This part focused on concerns related to fundamental rights, individual freedoms, the common good, the environment and the future of humanity. The issues surrounding the development of science and ethics are nothing new. There is indeed a very abundant literature exploring that dimension and the issues arising in that respect. It could be suggested to further complement this paper by integrating those reflections and make it clearer how the reflection we are facing is novel and differ from the issues encountered in other fields of knowledge.• In any case, the wish that an ethical reflection should go together with the development of AI and intelligent systems which should be designed with respect of fundamental rights, is something the CCBE strongly believes is desirable and necessary.	• This paper does not consider an essential component when applying fundamental rights and moral values, i.e. the possible conflicts and contradictions that may arise between them. We know that ethical problems arise when several fundamental principles are competing or in contradiction with each other. Pretending to build an AI in "compliance with Ethics" would presuppose that there will be a list of existing solutions available to solve all the possible conflicts. This is not the case. The CCBE considers that the paper omits developing these concerns and the inclusion of a note on the possible options how to deal with conflicting fundamental rights or moral values would be highly desirable. • Another important consideration concerns the question whether an AI system can be entrusted with the task of determining what is right or wrong in a given situation and to what extent we can consider to delegate ethical choices to an AI system? In this context, it is necessary to consider whether we accept the very principle of a delegation of powers to AI allowing it to settle a conflict between values or between the divergent interests and rights of several human beings. Would we accept that an AI determines solutions to ethical or moral problems, with the risk of causing harm because of its choices? This also relates to the question of liability and accountability of AI systems. • Technology and AI systems are reshaping the decision-making process in both public and private sectors and therefore have also the ability to reshape the relationship between decision-	Overall, having the capability to generate tremendous benefits for individuals and society, AI also gives rise to certain risks that should be properly managed. We must ensure to follow the road that maximises the benefits of AI while minimising its risks. A human-centric approach to AI should consider to keep in mind that the development and use of AI should not be seen as an end in itself, but as a means to increase human well-being. In this perspective, trustworthy AI is a qualitative factor.As noted above, the CCBE believes that specific consideration for the use and development of AI in the justice field in light of its potentialities should be separately developed in this paper. For example, the following important issues and questions may arise in this regard: • The judicial system is currently in charge of producing solutions to conflicts of norms. It is the role of the judicial system to provide individualised solutions to the ever-present conflicts between principles and ethical values. If the decision is made to let the AI develop its own solutions, then it is a transfer of a priori responsibility for moral choices from the judge to the machine. Here we have again the question of fundamental rights. Is it compatible with human dignity that machines judge men?• In practice, it would first and foremost be a transfer of responsibility for deciding, from the judge to the designer of the AI. Can such a transfer be considered if the ex-ante guarantees provided by the AI, at the design stage, are not of the same level as those existing in the	• The development of Artificial Intelligence, automation systems and other emerging technologies bring new challenges in terms of liability and data access and those issues should be carefully addressed and checked whether the current legal framework is adequate. In this regard, the CCBE is part of the Commission' Expert group on Liability and new technologies which aims is to provide the EU with expertise on the applicability of the Product Liability Directive to new technologies. • We would also like to bring attention to the HLEG that, from the point of view of legal practitioners, the CCBE suggested to the Council e-Justice Working Party the idea to establish a set of recommendations on the use of AI in the Justice field which has recently been included in the e-Justice Action Plan for the period 2019-2023.• The CCBE would welcome the possibility of contributing to this discussion on the issues around the use of AI applications and its possible impact from the point of view of Justice and legal practitioners. Due to the tight time schedule of the consultation, the comments submitted here are general remarks for the purpose of presenting a preliminary analysis of the issues set out in the Draft Ethics Guidelines for Trustworthy AI and are therefore still subject to any position being developed at a later stage by the CCBE on this topic.

non-technical suggestions. The work done is very sound, but it will remain quite largely theoretical unless a comprehensive case study assessment will complete it. We believe that the final version of the paper should include examples of implementation, on a case-by-case basis, in specific areas. We also note that among the four particular use cases of AI that the final version of the HLEG Guidelines will develop, the case of "Profiling and law enforcement" would be included. Moreover, the CCBE suggests to also include a case scenario on the use of AI in justice, for example in a trial proceeding. • The CCBE considers that there is a strong need for having a special discussion on the use of AI in the justice systems : As indicated in page 3, the authors acknowledge that, while the scope of those guidelines covers AI applications in general, different situations raise different challenges and thus, a tailored approach is needed given AI's context specificity. We understand that the scope of the paper is to provide general overarching principles, but by not considering the specifics of justice systems, we fear that some important issues around the use of AI and automation systems in the field of justice will not be tackled. Since justice plays such a large and special role in the society, we believe that this should be specifically discussed in this paper. In many cases the standards of human behaviour are being created (or at least directly applied) in the judicial process, the special use of AI in judicial systems should be more explored in this regard. This discussion could be added in the part 5 of Chapter I – Critical concerns raised by AI.In every way, the right to a fair trial is one of the fundamental rights, which is the basis for the "Trustworthy AI", and trustworthy justice is part of the rule of law and principle of democracy.

makers and citizens. Any conclusion in this respect should clearly state that it is not proven that the benefits of AI are greater than the risks. There can be some benefits from the perspective of public authorities to make use of such systems for reasons like efficiency or reduction of costs, or from the perspective of citizens: more impartiality (humans vs. machines); equality, legal certainty, and consistency through automated decision-making. However, ensuring that systems comply with the rule of law is not apparent and this aspect should be carefully considered, especially when the use of automated systems may endanger the principle of procedural rights. • Recent experience demonstrates that transformation of paper-based processes to electronic ones sometimes resulted in transferring administrative burdens to citizens. In the public sector, these transformations resulted in both cost savings (decrease in the administrative staff required) and faster, more effective processes on the government side. However, such transformations generated further costs on the citizen's side (including their representatives), e.g. new integration costs that were not present in paper-based processes, or IT security costs due to new threats. Citizens may have saved on the post € 10, but have to spend € 100 on IT security and updates. With regard to further transformation of such processes to be able to make use of the increased capabilities provided by AI, it is important to take a look at the context of the processes transformed, and to inspect any unintended effects such transformations may have. These unintended effects are often not cost related, but result in indirect loss of importance of certain values, as collateral damage, such as confidentiality of communications or privacy. We suggest that with regard to transformation of processes in the public sector using AI, this requirement of investigating the unintended effect on context should also be included.• Similar to the problem mentioned in the previous point, the requirements in Chapter II are often in contradiction with each other. E.g. transparency and safety often requires features that result in weakening privacy requirements. If someone designs a system for users with diverse and different disabilities, this might also result in unintentional discrimination for technical reasons. Even with the best intention and with an enthusiasm for transparency, a designer of an AI system could get into a conflict where the trained features show existing discrimination, and the designer has to intervene to avoid further strengthening the discrimination present in the society. We think it could be useful to mention either in Chapter II or Chapter III that in assessing a Trustworthy AI, one also has to address all the requirements identified (even if the list is not exhaustive), and make a human decision on priorities.• From the paper (see page 7 point 3.3), it is clear that the authors consider the use of AI systems in judicial systems in a way that allows AI to take decisions. Perhaps, this could be the case in some straightforward decision-making process when there is no doubt about applicable moral standards and the way they should be applied – giving the possibility of a human review. An example of a parking ticket violation can be provided in this context. This will be different in complex judicial system? Should the AI designer be independent and impartial, as the judge should be? At least in the event of an AI being implemented to assist (replace?) the judge, this seems to be a necessity.• Is it otherwise possible to be satisfied with a simple ex-post control, in the form of compensation for errors made by the AI, or a right of appeal from the AI's decisions?• The judicial context also gives rise to other specificities:- It can never be assumed that the AI systems implemented effectively respect the principles governing the functioning of AI. The right to review must in all cases be open to the parties concerned. - Lawyers should be given the opportunity to verify the compliance of the systems used with the principles identified.- Justice is an area in which transparency and accountability are particularly essential. This transparency extends to the design conditions and the identity of the system designers.These considerations also highlight the need to always carefully consider the role of AI in the decision-making process (e.g. as illustrated before, in certain situations AI systems should play a supportive role only).

					cases when values are to be applied differently or interpreted in a new way. In these cases, AI should play a supportive role only. This again underlines the need to more carefully consider the topic of the use of AI in the justice field. • Also, the paper seems to focus on AI-human relations and does not really consider the relations between two AI systems that can influence humans. • The human-centric conception is a leading principle of the paper. In this respect, it should be considered to state – to avoid any ambiguity – that the ability of artificial intelligence systems to make decisions autonomous from human control, requires specific attention, especially when these decisions can change the legal position of an individual or entity and imply making choices between concurring or conflicting values. In this respect, such autonomous process should be even more carefully assessed, from an ethical point of view, before being introduced in specific fields (such as Judiciary and legal services).• Humans are the key element in this paper. However, the CCBE wonders whether the effect of AI behaviour towards humanity (as something broader than a human) should not be explored as well? • The question of 'transparency' also encompasses the question of who 'owns' an algorithm, e.g. an individual developer, a multinational company with a wide-spread business model or even a state-owned enterprise. This aspect should be reflected more in the paper.• Whilst the problem of errors is recognised, a discussion on how to remedy such ' 'malfunctions' is missing.• It might be necessary to conduct a more in-depth analysis of the potential impact of AI with regard to the rights protected by the EU Charter of Fundamental Rights instead of only making reference to 'fundamental rights'.		
Michał	Zakrzewski	APPLiA - Home Appliance Europe	no comments	no comments	no comments	no comments	APPLIA proposed AI definition: Artificial intelligence (AI) refers to computer systems designed by humans that, given a complex task, act by processing the structured or unstructured data collected in their environment according to a set of instructions and operations, determining the best action(s) to take to perform the given task, via software or hardware actuators. AI computer systems can also adapt their behavior by analysing how the environment is affected by their previous actions.
Jens	Lidén	TCO Sweden		In this chapter, the expert group proposes that artificial intelligence (AI) should be developed, deployed and used with an "ethical purpose", grounded in fundamental rights, ethical principles and values. It is suggested that the fundamental rights commitment of the EU Treaties and Charter of Fundamental Rights should be used as a fundament to identify ethical principles for AI. It is in our view crucial that AI is developed and used in compliance with fundamental rights and international human rights law. This is a self-evident key aspect in the work to achieve trustworthy AI. For this reason, we are pleased to note that the draft guidelines is using a "rights-based approach"	This chapter is suggesting concrete requirements and methods to achieve and implement trustworthy AI. The suggested requirements and methods are grounded in the rights, principles and values mapped out in chapter I. Two of the most important elements in achieving trustworthy AI at work are, in our view, transparency and accountability. Workers who are affected by AI systems should be knowingly informed about the AI systems in operation at work, the purpose of the systems and the mechanisms by which the systems make decisions. All personal data gathered by the systems should also be used in compliance with the GDPR.	TCO welcomes the ambition of reflecting on four cases of AI. We suggest that one of these cases should involve implementation of AI systems at work. One issue among many involve generation of, ownership over and dissemination of AI input data through regular work activities (using work tools connected to a local network or the Internet).	AI is one of the most transformative forces of society today and AI will alter our conception of many things, including but not limited to the labour market. The labour market is however one of the societal sectors mostly exposed to AI and tech-led developments. For this reason, it is of utmost importance to ensure that AI is developed, deployed and used for ethical purposes at work. In addition, it needs to be ensured that AI is developed and used fairly and transparent for workers on the labour market affected by AI. Transparency, explainability and accountability are key elements in this respect. TCO supports the Expert Group's work to establish ethical guidelines for AI. TCO is

				in order to form an ethical framework for AI. We have, however, a few detailed comments on the draft in this respect. The fundamental rights commitment of EU law is a natural stepping stone when discussing ethics in an AI context. However, many member states are bound by additional commitments under international law in the field of fundamental rights, for example rights commitments under ILO and UN conventions and declarations. Our position is that these fundamental rights, based on ILO and UN legislation, cannot be ignored when forming an ethical framework for AI. Furthermore, the member states are bound by human rights commitments under national constitutions. For these reasons we request for the guidelines to refer specifically to such additional rights commitments and to elaborate on the legal relationship between these commitments and the commitments under EU law. Regarding the critical concerns section, section 5 of this chapter, we call for the wording "critical concerns" (title point 5 and 5.5) to be put as "red lines" as was the case in the initial draft. Furthermore, we wish for the guidelines to explicitly address the employer-worker relationship as a situation raising critical concerns.	It is of general importance that AI systems include mechanisms for accountability. In the event of a negative outcome, there has to be such mechanisms in place in order to hold those in control of the system accountable. However, from our perspective it is of specific importance that the systems are not in any way obstructing workers from holding the employer accountable under labour law. Further, it is crucial to recognise that "opt out-mechanisms" might not be sufficient for workers, as opting out could mean one has to opt out of a job altogether. TCO notes that transparency and accountability are mentioned as specific requirements to achieve trustworthy AI. However, TCO wishes for the guidelines to elaborate on these parts from a labour market perspective explicitly. Discussions on Data Governance would for example benefit from including situations in which workers generate AI input data through their regular work. The social partners and national trade unions can play a significant role in the work of achieving trustworthy AI. We believe that this aspect should be addressed more elaborately when discussing non-technical methods for ensuring ethical AI.		pleased to note that the draft guidelines have a rights-based approach and that the employer-worker relationship is described as a situation to be particularly aware about. However, in order to ensure ethical development and usage of AI at work we believe that the ethical guidelines must elaborate more precisely on the certain risks that AI rises on the labour market.
Rauni	Söderlund	Trade Union Pro	- UNI Europa ICTS welcomes the possibility to contribute to the stakeholders' consultation and underlines the importance of a broad public debate and information on AI. This debate must result in clear ethical and social guidelines and standards with the aim of improving the living and working conditions of European citizens.- We acknowledge the innovative potential of AI and new technologies that can be beneficial for our society. However, these new technologies also create challenges and we are concerned about the possible risks and consequences relating to working conditions, skills and training, ethics, equality, health and safety (among others). Therefore, UNI Europa would like to underline the importance of addressing AI technologies and robotization as topics for collective bargaining at all levels (company, national and European). AI and robotics have a huge impact on the future labour market, as jobs will sometimes disappear or be transformed and other jobs will be created. We need to accompany this process and address the question of skills and training for the future workforce: need to ensure that training on necessary digital skills is provided by education institutions and companies, and that it is not the sole responsibility of the worker to keep up with the rapid technological developments. Employability needs to be promoted through upskilling and reskilling schemes for workers. Investment in formal, informal and life-long learning is key; we must enable people to work with AI or invest in competences that AI will not cover. It is important to develop action plans at EU and national level together with education providers and social partners in order to modernize education and vocational training. We therefore welcome the call from the ILO Global Commission on the Future of Work for "a formal recognition of a universal	- UNI Europa supports the human-centric or human-in-command approach suggested in the guidelines. We agree that it is necessary that humans always need to remain in control of technology and machines. Likewise, we agree that the use of AI needs to respect European values and fundamental rights.- We recommend an expansion of 3.2 "In an AI context, freedom of the individual requires protection from direct or indirect coercion, surveillance, deception or manipulation." Much profiling that lies at the heart of AI systems relies on a degree of manipulation of data. This is not least relevant in relation to the world of work and especially in the use of AI in Human Resources. - We welcome that the HLEG understands the need to ensure that those involved in the development and marketing of AI (researchers, engineers, designers etc.) act in accordance with ethic and social responsibility criteria. This should be addressed by changing educational priorities for technical subjects and by providing lifelong learning opportunities (e.g. by incorporating ethics and the humanities into training courses in engineering). - Organisations and companies should develop tools to facilitate ethical discussions and decision-making throughout the whole design process. This should be completed by internal training programs on ethics for all employees. Such training should help employees understand the AI systems themselves, their rights in relation to said systems and their possibilities of redress, complaint etc. - AI should provide an opportunity for workers to apply their skills and competences to the fullest while at the same time remain owner of the production process. This includes the principle of transparency in the use of AI systems in HR, like the hiring of employees or the performance assessment of staff. It is	- In order to achieve "trustworthy AI", we need to establish public, independent and autonomous organisations that can control and audit (labour) algorithms (e.g. to identify underlying biases and the objectivity of data sets that train algorithms). Likewise, the implementation of the ethical guidelines on AI must be monitored. A European observatory focusing on the ethics in AI systems could play the role of an independent watchdog, including in business.- We would like the advice „to always keep record of the data that is fed to the AI systems" from the heading of data governance included under Accountability. For workers, it is paramount that the datasets used to evaluate performance, or in hiring or firing processes is transparent and can be accounted for. - The explanation of the principle of autonomy covers the question of AI at work only in a footnote, whereas this is an important issue that should be given a more prominent place. We would like to highlight the right of workers to individually and collectively opt out or withdraw from the use of AI systems (or a decision chosen by an AI system) if they undermine the workers' autonomy, decision making competence or disrespect fundamental rights and ethical principles. We recommend the inclusion of a special chapter that provides for ethical guidelines on AI in the work environment to address these issues more in detail.- UNI Europa ICTS welcomes that the HLEG on AI acknowledges the importance of social dialogue to realise trustworthy AI. We would like to add that the involvement of social partners, and in particular employee representatives, should not only take place regarding the general public debate on AI. Social partners should be involved in the establishment of codes of conducts, of standardisation schemes, development of training and in the proposed	- We welcome that processes shall be examined in order "to allow a human control, if needed" (assessment list – governing AI autonomy). In so doing it should not be a question "to keep a human in the loop". We need clearly defined measures, that empower people to exert this control in all processes – regarding resources (technical equipment etc), organisational needs (time, liability, etc.) and qualification.- We suggest extending the list on the assessment of use cases (p.28) and add the question of processes, in order to use AI to ensure decent work (development and impact assessment).	- UNI Europa ICTS welcomes the call for Accountability Governance on page 21. The establishment of Data/AI Governance Councils in companies will indeed strengthen the accountability of AI systems and will address a weakness in the GDPR. The Council should consist of shop stewards and management and be responsible for holding management accountable and transparent to the use of AI and data. Whistleblowers should be able to address concerns to the Council and mandate the council to investigate on reported issues. - We welcome the process of developing guidelines for a trustworthy AI made in Europe, which encompasses corresponding "guidelines made in Europe", but would like to raise the question why non-European companies such as Google were granted full membership and full participatory right in the High- Level Expert Group. The status of associate expert would be more appropriate.- UNI Europa also supports the position of the ETUC regarding this consultation.

			entitlement to lifelong learning and the establishment of an effective lifelong learning system. (https://www.ilo.org/global/topics/future-of-work/publications/WCMS_662410/lang--en/index.htm)- The social partners play a key role in this and the EU should cooperate with them and national governments in order to identify which job sectors will be affected by AI. We need to understand the timeline and extent of changes in the labour market. The involvement of social partners is a must to find appropriate and future-proof solutions to concerns relating to employment, training, the nature of work, (in)equality or social systems and collective bargaining, especially at sectoral level.- As AI and automation have the potential to transform not only simple tasks but very complex processes, we need to have a large public discussion about the areas in which the use of AI is reasonable and beneficial for society. Part of the debate should be the question of how the profits generated by AI should be re-invested for the common good by creating employment in domains such as care, health services, education or mobility. Employees should participate in the distribution of profits, e.g. through wage increases or reduction of working time. Moreover, AI wins should be used to strengthen social security systems. This could be a measure to address the problem of future job losses and the precarisation of employment relations in a-typical work (e.g. platform work) due to AI and automation.- It is therefore important to integrate the aspect of the quality of jobs, decent work and social progress into the ethical approach in order to create a balance with the purely economic objectives underlying the creation and use of AI and robotics. - The Human-centric approach (HCD) not only presupposes information, transparency, participation and traceability, but also requires specific negotiation processes regarding decision-making in view of the aims and implementation of AI-systems at a very early stage for stakeholders such as employees and their codetermination bodies.	important to safeguard the rights and freedoms of employees in line with non-discrimination rules as regards the processing of workers' data.- UNI Europa welcomes 5.1 – 5.4. We support that these examples raise real-life concerns of the adverse consequences of AI systems. - In 5.2. UNI Europa urges the group to expand on the issue of the human's right to know they are interacting with an AI identify. This could be done through a "labelling" system. For example, online bots should be labelled as such. Users should be made aware of the use of bots and AI in customer call-centre or help desks etc. - We would welcome that the employer-employee, employer-worker relation is explicitly mentioned in 5.3 as an example of power asymmetry. - Taking into account the power asymmetry in employer-worker relations, a separate point 3.6 on "workers' rights" should be added, which should contain the following points: "decent work by design", equal negotiation processes in the sense codetermination rights, informational self-determination of employees, non-discrimination principle and freedom of association including the right to strike. This is needed in order to secure worker's rights to co-decide on aims and application of AI systems, and create a legal framework.- Concerning the long-term risks and concerns we welcome that these should be considered. This could become an integral part of the accountability and audibility demands – i.e. that developers, users deployers etc need to reflect on the development/changing nature of the adopted AI as well as engage in predictions/forecasts of its future development scope and the consequences (positive and negative) hereof). - AI's influence does not only affect the world of work, but also democracy and society as a whole. We welcome that the draft refers to this point in Chapter I, paragraph 5.3., by stating that AI is not to be implemented in order to enable "citizen scoring" by a state/government. But this should also apply to private businesses. Neither states nor companies should be allowed nor have the possibility, to create human profiles such as "moral personality" or "ethical integrity". We reject the proposed opt-out-function and even possible "opt-in"-functions are not to be designed in a way that they conflict with fundamental human rights and possibly lead to the waiving of services that are useful for a person. AI-based services, that are important for work and life, must be designed in such a way that they do not require the collection of data which could be useable for human profiling.- Creating big data-bases always includes the risk of hackability as well as intentional and unintentional data-leaks. The guiding principle of "data-sovereignty" needs data-security in order to be viable. This implies explicitly not surveying data in areas that are of highly explosive nature for people in e.g. political, private or work-related areas. Fundamental rights as informational self-determination, the freedom of association and freedom of speech are not to be put at risk by creating such data-bases.	accountability governance. Employee participation and inclusion should take place early in the design, development and deployment of new technologies including AI and robotics. It is essential and important not only to inform and consult workers representatives in the work place or at branch level, but to enhance their co-determination rights and ensure their right to co-decide on the aims, reasons and implementation of AI at the workplace. - Social partners at all levels should be involved in the implementation at company, industry, national or international level, including through collective agreements setting standards. In this context, it is required to describe the negotiation processes, e.g. central control structures for sector-specific solutions (cf. ‚AI Now' Report 2018): „Governments need to regulate AI by expanding the powers of sector-specific agencies to oversee, audit, and monitor these technologies by domain." (https://ainowinstitute.org/AI_Now_2018_Report.pdf)- Regarding the principles of accountability and transparency, we need to establish mechanisms for the protection of whistle-blowers who disclose the risks of AI systems or the non-respect of ethical principles – especially in the case of employees in companies that develop such systems. Internal reporting of risks and violations should be supported and rules in place to ensure follow up. - Organisations and companies should pay attention to potential biases encoded in the system development, training data and model performance – especially those that my affect the most vulnerable. They could also establish an internal ethical review process to democratise the decision-making process- Companies should not only increase transparency regarding the design and development of AI systems, but also in organisational chains of responsibility.		
Natalia	Jaekel	Innenministerium Baden-Württemberg	Die Landesregierung Baden-Württemberg begrüßt den mit der Konsultation zu ethischen Leitlinien für Künstliche Intelligenz (KI) eingeleiteten Prozess für eine Debatte	Die Landesregierung Baden-Württemberg betont, dass im Rahmen der Debatte zu Grundrechten, Prinzipien und Werten das Primat der Menschenwürde stärker zu	Die Landesregierung Baden-Württemberg stimmt mit dem Entwurf darin überein, dass die angegebenen zehn Anforderungen für eine vertrauenswürdige KI (1.	Die Landesregierung Baden-Württemberg wird sich dafür einsetzen, dass das Land Baden-Württemberg und die von ihr geförderten Einrichtungen (insbesondere das	Die Landesregierung Baden-Württemberg bekennt sich zu einer menschenzentrierten Gestaltung der KI und dem Ziel einer Kultur der "Vertrauenswürdigen KI made in

			zur Vertrauenswürdigen KI made in Europe.	berücksichtigen ist. Dabei soll durch den Einsatz von KI der Mensch nicht prinzipiell ersetzt werden, vielmehr sollen Menschen und KI sich gegenseitig ergänzen, um deren jeweiligen besonderen Stärken zu nutzen und Schwächen zu kompensieren.	Verantwortlichkeit, 2. Datenqualität, 3. Konzeption für alle, 4. Menschliche Überwachung, 5. Nichtdiskriminierung, 6. Respekt und Verbesserung der menschlichen Selbstbestimmung, 7. Achtung der Privatsphäre, 8. Robustheit, 9. Sicherheit und 10. Transparenz) zu berücksichtigen sind. Allerdings wird die Einschätzung im Entwurf kritisch bewertet, dass diese zehn Kriterien gleich wichtig sein sollen. Im Hinblick auf die besondere Bedeutung der Menschenwürde kommt den Kriterien "Nichtdiskriminierung", "Respekt und Verbesserung der menschlichen Selbstbestimmung" sowie "Achtung der Privatsphäre" eine erhöhte Bedeutung zu.	Cyber Valley und der Digital Hub Artificial Intelligence in der Technologieregion Karlsruhe) an der Fortschreibung der offenen Liste für die Bewertung von vertrauenswürdiger KI mitwirken.	Europe". Sie wird im Dialog mit ihren Bürgerinnen und Bürgern daran mitwirken, dass in Baden-Württemberg die Bedingungen für weltweit führende Innovationen im Bereich der KI auf Basis dieser ethischen Leitlinien erfüllt werden. Dabei ist insbesondere die Menschenwürde als "value-by-design" zu beachten.
Anonymous	Anonymous	Anonymous		Referring to: 5.2 Covert AI systems A human always has to know if she/he is interacting with a human being or a machine, and it is the responsibility of AI developers and deployers that this is reliably achieved. Otherwise, people with the power to control AI are potentially able to manipulate humans on an unprecedented scale. AI developers and deployers should therefore ensure that humans are made aware of – or able to request and validate the fact that – they interact with an AI identity. Note that border-cases exist and complicate the matter – e.g. an AI-filtered voice spoken by a human. Androids can be considered covert AI systems, as they are robots that are built to be as human-like as possible. Their inclusion in human society might change our perception of humans and humanity. It should be born in mind that the confusion between humans and machines has multiple consequences such as attachment, influence, or reduction of the value of being human. [16] The development of humanoid and android robots should therefore undergo careful ethical assessment. I agree that a careful ethical assessment of the development of humanoid and android robots should be performed. However I would not restrict this assessment to humanoid and android robots but would argue that other robots (not humanoid looking ones) also evoke feelings and attachments with humans and are therefore just as worthy of a careful analysis. Humans develop relations to and identificate with technical systems no matter which form they have.			
Anonymous	Anonymous	Anonymous	No comments	3.4 Equality and non-discrimination (e.g. minorities): If an AI that does not work for a certain minority or ethnic group is by definition a violation. E.g. a face detector might work less well for certain groups or another example dermatologic screening might initially work on skin type 1-3, only when sufficient data becomes available it can be extended to other skin types. However, there is a correlation between skin type and certain ethnic groups, so the initial product will not roll out for all ethnic groups at once. 3.5 Citizens' rights: citizens should enjoy a right to be informed of any automated treatment of their data by government	1.5 Non-discrimination I would rather have a system that returns me the message that no reliable conclusion/action is available than one that gives me an erroneous result. And first training an AI for each possible minority would slow down to the level that it is not realistic. Furthermore could the release of an initial product to a group enable to gather more rare data, thus enabling the expansion of its user base in the long run. 2.1.4 Tracebility & auditability I understand the benefit of (internal or external) audits, but unclear what they would audit or what organizations currently do so.	The concern when a product is good enough to launch is not mentioned, whereas every system will make mistakes. Is a system good enough if its statistical decision error rate by the AI is better than statistical decision error rate by humans? Should we include accuracy numbers / clinical outcome numbers in product. When is it ethical to take the "collateral", also knowing that initial interaction will further improve the system.	The link between accountability and liability is nowhere mentioned, and perhaps not the task of the current task force, but at some level this discussion should be discussed. Concerning the transparency of the business model: if we're reselling information (=the results of analytics on the customer's data) to third parties, is it then good enough if we put that transparently in the EULA (End User License Agreement)

				bodies, and should have an opt-out option. What about traffic monitoring by government bodies,…? 4.3 The principle of autonomy: citizens should enjoy a right to be informed of any automated treatment of their data by government bodies, and should have an opt-out option. What about traffic monitoring by government bodies,…? A right to opt-out or be informed by AI, but in case of healthcare diagnoses: who ? a patient, doctor, operator or somebody in some video feed,..? Individuals or minorities are afforded similar opportunities as the status quo – this was mentioned here, but not sure if this would be the rule for non-discrimination, while it would be more realistic that they should not be hindered by an AI. 5.1 Identification without consent: Sometimes it is ethical, e.g. for detecting terrorist actions, …) but you sell a technology, not the use-case. And this could even differ from what government is working on the same use case. So it remains unclear what is ethical and what not in such cases.	2.2.3 Accountability Governance The mentioned idea of having a single person of contact/panel that checks the ethical aspects (or provides the needed processes) seems very relevant (similar to a data protection officer) both to ensure we are ethical, but also have a contact person where potential bias or risks could be reported. 2.2.5 Education and awareness To "educate" our users, can we make stats of our training data public? If there is a bias that we might be missing, we could also be warned.	
Cecilia	BONEFELD-DAHL	DIGITALEUROPE	Summary:DIGITALEUROPE supports the overall ambition and approach taken by the High Level Expert Group and the draft AI Ethics Guidelines. We support the goal of fostering the development and uptake of trustworthy AI, in the context of building a more competitive Europe. However, it is important to keep in mind the global market and the borderless nature of (digital) technology. In that sense, we question the term of "AI made in Europe". The Guidelines could further highlight more strongly the benefits and use of AI, for example in the manufacturing and industry sector, as well as potential for environment protection and economic growth.---Glossary:The definition for AI is not fully in line with the community of AI practitioners and the current state of art in AI technology. In particular, the statement that AI systems are "designed by humans" and are "deciding the best actions to take (according to pre-defined parameters) to achieve a given goal" appears outdated and ignores the existence of machine learning systems that are in fact not completely pre-defined by humans. The definition of bias does not reflect the actual scientific meaning in statistics, but instead overly focuses on the human element. It also mentions only with one sentence the potential for AI systems to support less biased decisions. This would be overstating the risks compared to the advantages of AI. For many AI applications, especially in the manufacturing and industrial sector, questions of bias and discrimination are much less relevant. Therefore such AI applications do not have any potential negative impact on end-users.The definition of trustworthy AI is clearer in the glossary section than the one used in the executive summary. In particular, it underlines that fundamental rights and regulations should be complied with during the development, deployment and use of AI. It is not the AI system itself that respects these rights and regulations.---Introduction:DIGITALEUROPE supports the recognition in the Guidelines that existing law and regulation in Europe already apply to AI. Therefore, there is no apparent legal vacuum. Consequently, AI	I. Respecting Fundamental Rights, Principles and Values - Ethical PurposeThroughout this part, the voluntary nature of the Guidelines is not properly reflected. The Guidelines read: "the section can be coined as governing the "ethical purpose" and it "identifies the requirements for trustworthy AI" .Yet, the Guidelines should be providing guidance.I.2. From Fundamental rights to Principles and ValuesInformed consent has been a valuable tool to empower citizens and give them control over data. But it always had a limited effect due to the burden it places on individuals to understand how data is collected, processed, used. The legitimate interest of the entity processing data should be balanced against the legitimate expectations of the individuals, so that it can supplement consent where context is appropriate. This would work in concert with substantial protections to individuals and obligations on organisations (e.g. accountability approaches). As the concept of informed consent traditionally belongs to the data protection sphere, DIGITALEUROPE suggests clarifying how this would apply to the ethical dimension, which includes privacy but is much broader in scope encompassing many more fundamental rights. I.3. Fundamental Rights of Human BeingsIn the section on "Human Dignity", we suggest utilising language to include "respect", without seeming to exclude more mundane applications. At the same time, it should uphold those principles (e.g. AI in entertainment):"To specify the development or application of AI in line with human dignity, one can further articulate that AI systems are developed in a manner which upholds serves and protects humans' physical and moral integrity, personal and cultural sense of identity as well as the satisfaction of their essential needs".In the section on "Citizen's rights", DIGITALEUROPE is concerned that the Guidelines propose measures that are at best unclear but often completely impracticable or at worst impossible to implement. For example, the proposal to "systematically be offered to express opt out" of "automated treatment of their data by government bodies" seems not	II.1.1. "Accountability"The accountability paragraph ignores the important role of accountability processes necessary within the organisation developing or deploying AI systems, i.e. precautionary measures on the one hand, as well as on the other hand policies and procedures which are always in place to address specific issues or incidents as they arise. The description also seems oriented more to liability rather than to accountability. DIGITALEUROPE would suggest the following wording instead for this paragraph: "Effective AI governance should include accountability measures, which could be very diverse in choice depending on the goals. Accountability can be described as the ability to demonstrate that appropriate measures have been put in place by an organization to minimise risks identified for the specific AI system and usage. These technical or organisational measures should be tailored based on each business' needs as well as the specific risks themselves. Consequently, regulators could deem accountability measures as a mitigating factor in case of incidents. […]"II.1.2. "Data governance"The data governance section omits best practices. It makes no mention of the traceability of data sources and data transformations, any documentation on the quality and nature of data etc. It also ignores the problem of re-identification of individuals following the combination of data sets. Furthermore, it wrongly assumes that "biases can be "prune[d] away before engaging in training". This may not always be possible and contradicts a later statement that underlines "data always carries some kind of bias."Suggestions such as "it is advisable to always keep record of the data that is fed to the AI systems" may in fact not always be compatible with EU data protection laws. It is also not clear what in practice is meant by: "To trust the data gathering process, it must be ensured that such data will not be used against the individuals who provided the data." DIGITALEUROPE additionally underlines the importance of data quality. Quality of the AI systems and solutions is deeply affected by data quality. Thus, the	This chapter needs to be further developed and supplemented with more analysis and insight. Its goal is to provide pragmatic guidance for organisations and businesses. It could also be linked better with existing regulation already, such as in particular the GDPR. 1. Accountability"Was a diversity and inclusiveness policy considered in relation to recruitment and retention of staff working on AI to ensure diversity of background?" We suggest moving this point to 'Design for all'.3. Design for all"Is the system equitable in use? " – This question needs to be re-phrased. It is a very high-level and not at all practical question to assess or answer.4. Governing AI autonomy"What measures have been taken to ensure that an AI system always makes decisions that are under the overall responsibility of human beings?" – This overlaps to some extent with Accountability. Also the question is only relevant if the developer and user are the same.6. Respect for privacy"How can users seek information about the use of their data valid consent and how can such consent be revoked?" – This needs to be adapted to align with GDPR and the legal bases for personal data processing other than consent.7. Respect for (& Enhancement of) Human Autonomy"Is the user informed in case of risks on human mental integrity (nudging) by the product?" This is a vague and unclear question on the undefined 'nudging'. This is practically not always possible to assess or communicate.10. Transparency"What measures are put in place to inform on the product's accuracy? On the reasons/criteria behind outcomes of the product?" This should be rephrased. It should address how to define accuracy and whether this is always relevant."Is the nature of the product or technology, and the potential risks or perceived risks (e.g. around biases) thereof, communicated in a way that the intended users, third parties and the general public can access and understand?" This repeats content of the 'purpose' section. It is also a very broad requirement, i.e. how is it envisaged that a developer could communicate to the 'general public' about all these very detailed	DIGITALEUROPE's main concern is that the overall tone of the Guidelines, at this time and in this version of the draft document, is too negative. Positive aspects of AI for society are not addressed with the same emphasis. The guidance is more about "what not to do", instead of "what to do". The Guidelines should therefore be aligned more closely to existing processes, better reflect the point that a technology-neutral, European regulations already safeguard many (if not all) of the mentioned points and acknowledge that AI has been used for a long time already in certain sectors. We should further aim to build a framework that is target-oriented, future-proof and consistently coordinated (and avoid a proliferation of divergent or contradictory measures).We hope this input will contribute towards a more balanced outcome towards which we will continue to strive through our participation in the HLEG.DIGITALEUROPE will continue to take a constructive approach to help deliver clear and actionable Guidelines. We will continue to contribute to the Guidelines development with particular focus on the assessment list and use cases, recognising the different types of AI implementation and context of AI usage and deployment.

Ethics Guidelines should have the important role of assessing the use of AI in specific contexts and situations. The Guidelines are key to acknowledging that protecting individuals and their data goes beyond legal compliance requirements: it means embracing societal values and working to build a much-needed trust in technologies and their impact on people.We also support the inclusion of governments and regulatory bodies in the set of addressees and stakeholders. Public institutions are valuable developers and users of AI.DIGITALEUROPE recommends that the proposed endorsement process is further discussed and re-assessed. In practice, an endorsement holds potential legal consequences, which would be at odds with the voluntary and non-binding nature of the Guidelines.

to be compatible with current practices (e.g. in healthcare, taxation). The emphasis should rather be on ensuring adequate technical safeguards (such as de-identification techniques and strong encryption) as well as a sound legal basis to institutionalise those automated practices in specific and well-identified contexts. Overall, the text appears to have a bias against AI expressed in rather negative statements even where this is not necessary. For example, rather than reading “AI systems must not interfere with democratic processes or undermine the plurality of values and life choices central to a democratic society”, the text could read: “AI systems should serve to further democratic processes and the plurality of values and life choices central to a democratic society.” Further, AI systems do not only “hold the potential to improve the scale and efficiency of government […] services” but they “are already” improving them.I.4. Ethical PrinciplesThere is a possibility that different principles may conflict in practice. There may be a need to examine the potential trade-offs in implementing these principles. Further, it should be cautioned against the assumption that everything can be addressed by an ‘ethical expert’. In fact, there will be a need to gather expertise from various sources, including legal and sector-specific.On the principle of “Do No Harm”, the paragraph should be phrased to focus less specifically so on data collection and profiling, which is also not always relevant in many AI use cases.DIGITALEUROPE would also suggest the following change to footnote 12 on environmental awareness: “Items to consider here are the positive and negative environmental impacts of the large amounts of computing power to run AI systems and the application of voluntary Data Centre initiatives such as the EU Code of Conduct to optimise operation within these facilities, the data warehouses needed for storage of data, and the procurement of minerals to fuel the batteries needed for all devices involved in an AI system. For the latter, these minerals most often come from a mine without certification in an under-developed country and contribute to the inhumane treatment of individuals.”On the principle of “Preserve Human Agency”, DIGITALEUROPE finds that a general right to opt-out or withdrawal may in practice be impossible, or also unnecessary or cause a harm to others. It is crucial to take the specific context and use case into account.In more detail, a right to decide to be subject (or not) to AI, a right to opt out and a right to withdraw significantly reduces the possibility to make effective use of AI systems. AI relies on large volumes of retrospective data, making the execution of these rights impossible for any AI system, especially since, typically, AI systems will further use the input by users to improve the algorithms the AI system is built of. These rights are relevant in the context of the GDPR, which regulates data protection and fully applies when personal data is processed by AI systems. But these principles cannot be merely extrapolated in the context of AI systems, which are not limited to personal data processing. On the principle of “Be Fair”, DIGITIALEUROPE argues that the term ‘effective redress’, as introduced in this section, should not be presented as a quality derived from the fact that an AI system is or not in place. It is rather a quality derived

quality of the datasets and knowledge on the analysis of bias and other data-related issues are vital to this European project. As noted earlier, many questions regarding potential bias or discrimination are not linked to the analysis methodology or AI algorithm as such, but rather to the input data provided.II.1.3. “Design for all”The sentence "Systems should be designed in a way that allows all citizens to use the products or services, regardless of their age, disability status or social status" should be changed to "Systems should be designed in a way that considers usability and accessibility so that the products or services should be inclusive and can be accepted by as many citizens as possible, regardless of their age, disability status or social status." Some systems are designed for specific users and applications. For instance, systems could be developed for a specific manufacturing process or for medical professionals and for employees that have a particular set of skills or expertise. In the industrial sector, this concept is much less relevant or applicable.Further, DIGITALEUROPE cautions against measures that would risk putting excessive costs on developers. Especially smaller developers, without resources to incorporate these exaggerated functions into their products and/or programmes, will be negatively impacted by these measures. Considering how fragile the nascent European industry is as well as the efforts to facilitate the uptake of the AI industry, it seems counter-intuitive to burden entrepreneurs and developers with more requirements. Therefore, we would like to at least clarify that this should not be a legal requirement but more of a recommendation to businesses.II.1.6. Respect for (& Enhancement of) Human AutonomyDIGITALEUROPE advises that this section needs further elaboration to offer more clarity and become more concrete. Terminology on ‘nudging’ and ‘extreme’ are not defined and cannot necessarily be assessed by the addressees of the Guidelines. The section also seems overly focused towards AI used in a B2C context and online shopping or personalisation. This seems to be an oversimplification of how personalisation works, noting also that it can be used to augment human autonomy, for instance by analysing more complex texts to distil and provide more useful information to the user. II.1.8. “Robustness”This section is still too vague and misses important elements such as being transparent on the level of confidence with which predictions are made or the level of uncertainty respectively. This was in an earlier version of the Guidelines and it is unclear why it was dropped.II.1.10 “Transparency”The term “development processes” should be clarified. The term should not refer to the design process for system software., The process that a company follows while designing its system software is one of the important factors that differentiate the company in question from its competitors, and thus disclosing these types of processes would raise issues around trade secrets and intellectual property protection.Moreover, the “transparency” section overly focuses on the perception of looking into the “black box”. Given that this type of transparency may not always be possible due to the complexity of systems or their nature (e.g. self-learning systems), it is important to

questions?

from a legal right or judicial procedure applied to a wrong, independently of the technology used. In other words, AI systems should not add to or subtract from the redress rights stemming from the implementation of the law and judiciary proceedings. At the same time, defining multiple avenues to guarantee to all citizens' access to judicial redress would be paramount.Furthermore, the phrasing of this section would need to be clarified, in order not to open the door for an increased amount of uncertainty, as businesses may risk having to compensate someone based on data practices that "may no longer be aligned with human beings individual or collective preferences". Our understanding is it either implies risk of unreasonable costs for the industry or needs to clarify in which circumstances a business may be liable for "inflicting harm on users" based on this unclear definition.On the principle of "Operate Transparently", DIGITALEUROPE does not support the following point: "Business model transparency means that human beings are knowingly informed of the intention of developers and technology implementers of AI systems." This is not a question pertaining to AI and in no other field do we require 'business model transparency'. Further, while DIGITALEUROPE fully agrees that measured transparency is indeed a key element in creating trust in AI systems, the Guidelines again propose impracticable measures when they read "AI systems [should be] intelligible by human beings at varying levels of comprehension and expertise." This would essentially put a limit to innovation and allow only simplistic systems and AI models.Similarly, "Individuals and groups may request evidence of the baseline parameters and instructions given as inputs for AI decision making" may technically not be possible (e.g. self-learning systems) or reveal competitively relevant information and intellectual property.I.5. Critical concerns raised by AIDIGITALEUROPE finds that some statements do not reflect the current and foreseeable state of technology. For example "Androids can be considered covert AI systems, as they are robots that are built to be as human-like as possible. Their inclusion in human society might change our perception of humans and humanity. […]".DIGITALEUROPE therefore suggests deleting the section on longer-term concerns, given that is still very speculative and lacks any practical relevance. There is no apparent way to envisage all possible scenarios. If these Guidelines are to be practical and immediately applicable, the focus should remain on what is currently available. The principles need to be broad enough to inform decisions on scenarios we cannot foresee today.

focus on the input and, even more, on the output stage to foster transparency. Particularly, the latter seems to be missing.Further, the wording in the following statement is too broad: "for uses that can have from all models that use human data or affect human beings or can have other morally significant impact." Does it imply full transparency for anything using "human data" (how would this data be defined?) in any way? We suggest introducing some nuance. Not all uses of AI require the same level of scrutiny or "transparency". As also mentioned earlier in the Guidelines "different situations raise different challenges", and the application of the principles should be context-specific. The wording could therefore be adapted to focus on AI systems involving high-stake decisions (those having legal effects or negatively affecting human beings). We also see an important role here for governments to consider which AI implementations require higher degrees of transparency and explainability to mitigate discrimination and harm to individuals.II.2.1. Technical methodsDIGITALEUROPE suggests changing the phrase: "be able to take adversarial data and attacks into account" to "be able to take predictable adversarial data and attacks into account".Further, regarding the technical measures, they can still acknowledge the potential need for human intervention to explain complex issues. We support in that sense the following point: "The development of human- machine interfaces that provide mechanisms for understanding the system's behaviour can assist in this regard."The "traceability & auditability" section appears to hold AI systems to higher standards than is currently the case with decisions taken by humans. Also, there is no real definition of what is meant by "transparent" and "understandable", which is key to having practicable guidance.The "standardisation" section also appears to overly focus on standardising the design of AI systems, rather than APIs and interfaces. It is unclear what is meant by this standardisation.

Ad	Kok	Actuarial Association of Europe	ACTUARIAL ASSOCIATION OF EUROPEFEEDBACK TO THE HIGH-LEVEL EXPERT GROUP THE DRAFT ETHICS GUIDELINES FOR TRUSTWORTHY AIDear Mrs. Smuha,Thank you for giving us the opportunity to react on the Draft Ethics Guidelines for Trustworthy AI by the AI HLEG.We have read the draft ethics guidelines with big interest and we appreciate the quality of the document very much. The discussion about trustworthy AI is, for our profession, of utmost importance. Firstly, because our activities are affected and secondly because we can enrich the discussion with our high-level experience – especially in the financial world – and the long history of our Profession.As it is always difficult to define ethical principles for an entire scientific discipline, the Actuarial Profession chose an approach – for its professionals – which focuses on the individual. For example, if you take the discipline of statistics – which is very close to the discipline of AI –, it is difficult to regulate the effects which can be achieved by doing statistics incorrectly, presenting statistical results in an inappropriate way or interpreting statistics wrongly (either by purpose or by lack of knowledge). Therefore, stakeholders of statistics as well as for AI have to address those challenges directly and in an appropriate manner. We would therefore prefer to analyse the different groups of stakeholders acting in the AI field in more depth and try to address the challenges that they are facing with concrete guidelines. One of these stakeholders are certainly the modelers (designing, coding, testing/validating) which should have similar tasks and responsibilities as actuaries have. As an example, to fulfil our roles, the Actuarial Profession has established four tools:1. Education standards: to become an Actuary, an education program has to be passed which has worldwide common minimum requirements and has more advanced requirements in Europe.2. Continuous Professional Development: Actuaries have to ensure that they are up-to-date in the fields which are relevant to the Profession as the world is moving forward.3. Code of conduct: all Actuaries in Europe follow the same code of conduct which is based on 5 principles: integrity, competence, compliance, impartiality and communication.4. Standards: with technical standards (European Standards of Actuarial Practice (ESAP), International Standards of Actuarial Practice (ISAP)), we ensure that actuaries produce outcomes of high quality following the principles which are stated in the code of conduct.We would appreciate to discuss our ideas further and explain in more depth how we believe the development of AI can be kept under control either via written communications but preferably in personal meetings with (parts of) your expert group. We are looking forward to future conversation on this important subject.Esko KivisaariAAE ChairpersonA brief description of Actuaries, Actuarial Science and the European Actuarial AssociationActuaries are mathematicians mainly based in the insurance and pensions industry but also in banking and more and more also in other industries. They are mainly responsible for calculating insurance rating schemes, for analysing financial positions, for evaluating required risk capital and provisions. In doing this, actuaries follow the principle of

“fairness” to act as an objective intermediate between the financial industry and the customers. The actuarial system is therefore an important part of financial regulation not only in Europe but all over the world.If you go back in the history to the roots of statistics, you will find some of the first applications in the field which nowadays are seen as actuarial. Therefore, a lot of developments and improvements in statistical analysis go back to activities and research done by Actuaries. From that, it is obvious that modern advanced statistical techniques like clustering, neural networks, etc. are already in the actuarial toolboxes for quite a while. Actuaries need statistical techniques mainly for two basic tasks: to assess differences on one hand and to assess similarities on the other hand between objects, cash-flow-streams, risks and even persons. Actuaries need to know what are the reasons and factors that trigger financial demand for individuals and how these demands can be financed by a large group of people. Fairness for both, the individual as well as the collective, is our guideline for our technical work.Actuarial Science is the discipline that applies mathematical and statistical methods to assess risk in insurance, finance and other industries and professions. In many countries, Actuaries must demonstrate their competence by passing a series of rigorous professional examinations. Actuarial Science includes a number of interrelated subjects, including mathematics, probability theory, statistics, finance, economics and computer science. Historically, Actuarial Science used deterministic models in the construction of tables and premiums. The science has gone through revolutionary changes since the 1980s due to the proliferation of high speed computers and the union of stochastic actuarial models with modern financial theory. While actuaries’ tasks are technical, it is important that they can explain and communicate what they are doing to a wider audience. In fact, the products that actuaries deliver are used by decision makers who are not necessarily specialized in Actuarial Science. Therefore, actuaries need to be able to explain their product/results in a transparent and clear manner to ensure that decisions are taken on a sound basis.As a consequence, the Actuarial Profession is based on a common understanding not only on what they should do but also on how they should fulfil their responsibilities technically and ethically. A common education system and a common ethic code of conduct are therefore the pillars on which our profession is built on.The Actuarial Association of Europe (AAE) was established in 1978 under the name Groupe Consultatif to represent actuarial associations in Europe. Its primary purpose is to provide advice and opinions to the various organisations of the European Union - the Commission, the Council of Ministers, the European Parliament, the European Supervisors and their committees – on actuarial issues in European legislation. The AAE currently has 36 member associations in 35 European countries, representing over 24,000 actuaries. Advice and comments provided by the AAE on behalf of the European actuarial profession are totally independent of industry interests.The Actuarial Association of Europe is registered in the EU Transparency Register under number 550855911144-54

Heritiana	Ranaivoson	MediaRoad (EU H2020 project)	The Guidelines seem to be addressing everyone. It appears that it is aimed at developers and deployers of AI solutions. This should be made more explicit. Two other audiences would deserve similar (but of course adapted) guidelines: (i) policy-makers and persons in charge of regulating AI; (ii) non-professional users of AI systems, i.e. general public. Category (i) may be the primary target for the next document to be drafted by the AI HLEG. Category (ii) would deserve their own document, which would probably give a greater place to data literacy, e.g. in relation to these Ethics GuidelinesThe Guidelines state that a "mechanism will be put in place that enables all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis." Hence the guidelines are not legally binding. We are curious to see how that will be done, and in particular what will be the incentives for AI developers and deployers to endorse these guidelines. We are afraid there is no way to ensure that once endorsed, these guidelines are indeed respected and followed-up.Conversely, regulation can play an important role in ensuring the principles are enforced. For example, the GDPR is mentioned in these Guidelines as a law to be respected by developers and deployers of AI. Beyond, we have high hopes for the GDPR as a tool to contribute to transparency (e.g. point 2 of Article 22). The GDPR obligation for data controllers demands that where personal data are being processed automatically (profiling), that the data subject is informed about the rationale of the processing and its possible consequences. This obligation will allow further scrutiny (for users, researchers, etc.) of content recommendation algorithms that are dominant in social media and search results.	Section 3.4 on "Equality, non-discrimination and solidarity including the rights of persons belonging to minorities" could be improved by specifying that in an AI context, services should be designed to be truly inclusive and accessible to all, independent notably of age and disabilities. Besides, minorities should be considered and included not only regarding access but also production. The latter (inclusion in production) is mentioned in Chapter 2, section 2, as a non-technical method to achieve requirements but we think it should be a principle (in this Chapter) or at least a Requirement (Chapter 2, section 1).The Principle of Non maleficence should include the idea that an AI developer should adopt a "data minimalism" approach, i.e. only asking for data they really need. The current text takes it as granted that it is a must to collect, store, use, etc. data and that what matters is the way it is done. On the contrary, it is important to always have as a question in the whole development and deployment process: is it even needed to collect; store, use, etc. data? This is in particular important for vulnerable demographics. Otherwise, we run the risk of increasing defiance against AI.In the same Principle, it is unclear why immigrants are put in the same category as children.Finally, also the same Principle mentions diversity and inclusion as principles, but it is equally important to have minorities also involved among developers and deployers, not only as users. This is however addressed in Chapter 2, section 2.The Principle of Explicability should start with the necessity of informing users that AI is being used, even before explaining. This is addressed only in chapter 3 (p.26).	It is good that diversity in data is mentioned as important for traceability (p.20). However, it needs to go beyond it applied to the dataset to be applied to the team developing and/or deploying. While the idea appears as one of the non-technical methods, diversity as a whole (hence not only regarding data but also teams, provided content, options, etc.) would deserve to be a requirement in this Chapter or a Principle in Chapter 1.	We advise that media becomes one of the use cases. It is key to closely consider how AI is transforming the media value chain, from content production to the audiences' experience. Account should also to be taken of abusive practices by online platforms involving a content recommendation to users. Recent scandals, such as Facebook and Cambridge Analytica, have raised debates around the potential impact of algorithms on elections and on the shaping of social movements. If there is any reality in phenomena such as filter bubbles or echo chambers; if AI plays a role in the distribution and spreading of fake news; more generally if diversity is really a core value for the European Union; it means that AI in the media sector already has an impact on our democracies, and is, therefore, a core issue.	The Draft Ethics Guidelines for Trustworthy AI are an important document. The major point is that it goes beyond a list of ethical principles and shows clear concern for the implementation. A particularly important Requirement is transparency, on how algorithms work, on the data they use, allowing users to understand the underlying biases.Our main general comment is that media is an important field to consider when drafting guidelines on AI, while it is currently belittled, e.g. on page 3 with the example of recommending a song (thus belittling the importance of music recommendation).New technologies, from smartphones voice-controlled speakers to wearable devices, are vastly increasing the amount of digital data we produce. In this context, AI is transforming the way media professionals analyze and transform data, with an impact on the whole society. For example, robot journalism (or news automation), while having started in the late 80s, is becoming an important part of news production (notably on sport or stock exchange). It speeds up news production and generates a vast amount of content in a matter of sector to be distributed and consumed in print and online. AI is also core in the automated personalization processes. Faced with content overload, consumers are supplied with recommendation systems designed to help them select what they are going to watch or listen to. AI-based recommendation systems are used to create tailored services, which are then pushed to mobile or web applications. AI has obvious benefits. It can play a key role in the standardization of solutions for accessibility services (e.g. for the semi-automatic generation of subtitles) as well as the application of new production methods. Robot journalism can free the time of journalists from doing a mundane task and give them more time to investigative journalism. It can easily adapt to human request and improve their reporting and can produce content in different languages, such as collecting daily economic data and writing similar articles based on the data every day.However, AI also raises challenges, which could prove dangerous for the media sectors, and beyond for the whole society. This could first represent a threat with the possibility of job losses for media workers, in particular journalists, replaced by news automation. The development of AI will also lead to the creation of new jobs. It is anyway at this stage difficult to predict the exact impact, whether positive or negative, on media workers.AI is also a technology used in the development of so-called deepfakes (manipulated digital videos that overlay another person's face onto a body or change what people actually said), making the lines between the fake and the real become increasingly blurred.Finally, regarding the impact of personalization, there is a risk of filter bubbles developing, that is to say, situations where users do not obtain access to and, hence, remain unaware about some types of content. Data-driven and fully automated personalization models are not sufficiently looking into how to include diversity and serendipity in algorithmic functions to broaden the consumer's experience.One common feature of these threats caused by AI is that the solution often relies on the AI technology itself, provided that it follows ethical guidelines, such as the ones drafted here. Thus, AI can

						be used to develop fact-checking tools that can prevent fake news to spread. For example, Truly Media is an online verification platform to authenticate content published online. The development of such tools itself requires careful consideration of what fake news are and more generally of the ethical rules that should frame their use.
Takao	Nitta	Cabinet Office of Japan		[Ethical Purpose] In this draft EU guidelines, a human-centric approach is upheld as the common thread in EU. Japan shares similar values, therefore we can support this draft guidelines in this regard.	[Requirements of Trustworthy AI 1 : Accountability] [Requirements of Trustworthy AI 5 : Non-Discrimination] [Requirements of Trustworthy AI 10 : Transparency] In Japan's AI social principles, we also recognize the importance of the "Accountability", "Non-Discrimination", and "Transparency". Therefore we support those principles. [Requirements of Trustworthy AI 2 : Data Governance] Regarding requirements 2 "Data Governance", Japan is also aware of ensuring trust of AI products, services and their supporting data. From the viewpoint of developing an environment to ensure the safety and security in the data utilization across national borders, we request to advance international cooperation with Japan in order to establish an environment for ensuring trusted AI and data. From the viewpoint of data integrity, we agree with the statement that it is advisable to keep record of the data, however we understand that it is not realistic to keep all records. Therefore we suggest that such kind of records to be kept should be limited to critical application only. [Requirements of Trustworthy AI 4 : Governance of AI Autonomy] We believe that it is important to set the level of governance according to the level of AI autonomy and the scope of AI application. However, especially in case of the governance of autonomous AI by governmental agencies, we believe that it is necessary to ensure the free business activities by private companies. In addition, it is desirable that international fora for the discussion on the governance of AI will be held. [Requirements of Trustworthy AI 7 : Respect for Privacy] In Japan, the privacy protection of personal data is also recognized as one of the important principles. It is consistent with the mind of EU guidelines. In Japan, since personal data has wide range from data that is greatly affected when	

					unjustly used (such as ideological beliefs, medical history etc) to the data in the public domain. We therefore believe that the balance between its utilization and protection needs to be carefully examined taking account of cultural backgrounds as well as common understanding by the society. [Technical and Non-Technical Methods to achieve Trustworthy AI] Regarding technical and non-technical methods to achieve Trustworthy AI, we believe basic concept on the selection of these two methods should be needed. For example, technical methods should prioritize than non-technical methods such as legal regulation, because the advance of AI technologies is very rapid. We appreciate to be able to cooperate with EU in order to organize the concept of the selection of methods. In addition, for the concrete arrangement of technical and non-technical methods to achieve Trustworthy AI, it is important to promote close collaboration between Japan and Europe, from the viewpoint of establishing a wide market including Japan and Europe.		
Véronique	De Keyser	European Humanist Federation	The introduction of the working document prepared by the HLEG AI sets out the basis for the construction of the next three chapters devoted to the ethical purpose, the realization of "trustworthy" AI and finally its assessment.It is therefore important that all the assumptions, principles and objectives on which the rest of the document is built are properly identified. In this respect, the European Humanist Federation feels that the following key elements are missing from this methodological basis or the consideration given to them is too shallow. This echoes throughout the rest of the document. == Benefits vs. Risks ==The document takes as an axiom the fact that "on the whole, AI's benefits outweigh its risks" without providing convincing evidence to back up this claim. As humanists, we are committed to technological progress and we measure the extent of the economic and strategic potential of AI. However, the understandable race to reap the benefits of this very potential should not result in a lack of methodological rigor, especially when the matter at hand is to draft ethical guidelines. == Social acceptance of risk ==In highly formalized administrative systems such as financial loan decisions, web searches, online customer services, personalized marketing based on social data, financial speculation, etc. intelligent tools are booming and provide outstanding results. If the algorithm used in a specific application respects what would have been a human decision and if the database that is used to train it is sufficiently exhaustive, there is in principle no bad surprise. However, even if there cannot be any guarantee of control in the design phase, it has to be possible to correct potential biases or inconsistencies post hoc.The document rightly suggests that the EU has to find a "road that maximises the benefits of AI while minimising its risks [and that] to ensure that we stay on the right track, a human-centric approach to AI is needed." It however fails to clearly recognize and acknowledge that risk zero does not exist. This in turn means that we have to ask	Chapter I provides and overview of the principles, rights and values that an ethical approach to AI should entail. While the structure of the chapter seems appropriate, key elements seem to be missing from the overall reasoning. Certain sections also need to be further refined.== Informed consent and societal control ==As it was the case in the introduction, this chapter as well considers users and citizens as mostly passive actors of the development of AI systems. Section 2 takes "informed consent" as the basis for operating trustworthy AI whereby people are to be "given enough information". As humanists, we think that citizens should be much more empowered to become actors of the development of AI. On the one hand, it is troubling that the acknowledgement that current practices - that clearly show that end users give consent without consideration despite being informed - is part of the section where the HLEG AI seems not to have reached consensus. On the other hand, even if there was consensus within the HLEG AI, given the impact and the pervasiveness of AI technologies, mere consent is not enough, even if it is "informed". End users, citizens, workers and society as a whole have to have a much more active role in the entire life cycle of AI technologies: from design to usage, including ex ante but also post hoc validation. The relationship between users and developers has to be bidirectional and continuous.This first of all entails that the concern about education should be central to the question of the future of AI. Without proper education citizens will not be able to reap the benefits of AI and minimize the risk of its usage. Their emancipation and their free will could be severely hampered, without them even realizing it. Current debates relating to the impact of social media on election outcomes clearly demonstrates this. However, beyond education, public debate about AI should be actively fostered by systematic societal control and oversight. This is why we propose the creation of an EU Observatory of	The fact that Chapter 1 left aside a number of issues or did not given them strong enough consideration results in these elements not being addressed enough throughout Chapter II.While the fact that the list of requirements discussed is acknowledged by the paper itself as non-exhaustive is on the one hand laudable – as indeed it lacks key elements – it is difficult to see how the current list – even if enhanced – will not become some kind of baseline in the future. It has to be clear from the outset that the nature of AI and the necessarily yet unknown applications and services that it will bring about carry the fundamental need that this set of "requirements" be continuously reviewed and submitted to societal oversight and validation. The document rightly recognizes this. However, the way this will be guaranteed and the way its current contents will be debated in the wider society are not clear.Creating a European Observatory of AI applications and services, as proposed by the European Humanist Federation, would definitely be a strong signal pointing in this direction. == Accountability, autonomy, safety and transparency ==In order for accountability to be implemented in practice, three elements are of utmost importance. First, users have to be provided with the tools to detect and understand anomalies and dysfunctions. This however presupposes that explainability issues are properly addressed. Second, procedures should be in place, allowing them to lodge complaints to specialized bodies creating a level playing field between them and the legal departments of private companies developing AI. Finally, thorough legal research has to prove that AI's specificity does not create loopholes that could be exploited at the expense of the consumer or user.It appears from the section on explainable AI research (XAI) that explanability of AI systems has not yet reached a satisfactory level, not by a longshot. In turn, this means that a number of requirements expressed in this chapter	Concerning Chapter III, the EHF does see the merits of the approach taken. It also welcomes the ambition of creating a number of use-case-specific sets of assessment questions. However, it warns that these lists – as the document itself acknowledges – are by definition incomplete. Here as well, because of AI's pervasiveness and the diversity of its applications, ex ante measures have to be complemented with systematic post hoc procedures including the design and management of feedback systems allowing to flag incidents.	The EHF welcomes the work done by the HLEG AI and is looking forward to see the results of this consultation included in the final version. However, we are also worried that despite what the document claims – that it is to become a living document – the final version will be used as a baseline to consider whether a specific AI application is deemed ethical by European standards – whether it is "trustworthy AI, made in Europe." The EHF also expresses its concerns that the understandable race for reaping the benefits of AI have resulted, within the HLEG AI in a certain lack of methodological rigor. For instance, the claim that AI's benefits largely outweigh its challenges is not proven. The declaration that there is no legal vacuum when it comes to AI seems very hasty. Despite the fact that the document acknowledges that the explicability of AI is by far not guaranteed, it does build some of its reasoning on this concept. The growing opacity of AI technologies and their extension to multiple domains of life pose less the problem of control over design or use - these have become almost impossible in some instances – than that of social impact and possible recourse in case of problems.In this sense, the working document is not realistic enough. It sets itself the goal to guarantee "trustworthy AI" without acknowledging the fact that maybe, to some extent, this can only be an aspiration. It focuses primarily on ex ante measures to minimize the risks of AI – and this is laudable. It however overlooks the importance of systematized feedback about the dysfunctions, threats and risks experienced by users. In disregarding the importance of societal control and validation, it hinders efficient detection of yet unknown threats and potentially undermines societal acceptance of AI, including the acceptance of the inherent risks that their usage might entail.This is why, to complement the ex ante measures listed, the EHF's main proposal concerns the creation of a European Observatory of AI Technologies and Services in charge of implementing social control at

a second fundamental question: What level of risk are we willing to accept, socially speaking? == Trustworthy AI vs. societal validation of AI ==The answer to the above question can only be given via democratic processes. It therefore becomes clear that many more efforts have to be invested in educating society about the risks represented by AI technologies. This becomes all the more critical since citizens are heavily impacted by the use of AI but are often not even conscious of the fact that other choices were theoretically possible.This dimension is addressed to some extent in the document but only marginally, despite the fact that it is absolutely central as it will define the level to which society will, on the long run, trust AI technologies.Beyond education – and this is also acknowledged in the document to some extent – end users should be involved at all levels of the design of AI services: from conception and design (ex-ante validation) to feedback after usage (post hoc validation and recourse).Post hoc mechanisms should not only be put in place within design teams at the discretion of AI developers. On the contrary, they should be systematic. This would result in strengthened public debate and informed societal oversight. Therefore, we would propose to add a third component to the definition of Trustworthy AI: "it should ensure an ethical purpose, it should be robust and should be socially controlled on an ongoing basis."We therefore propose the creation of a European Observatory of AI Technologies and Services in charge of implementing social control at any stage of the design, deployment or use, including post hoc end-user return of incidents.In this sense, AI technologies would not only be trustworthy, they would actually be trusted. An EU observatory would also address two other elements that the document rightfully captures: given the nature and pervasiveness of AI technologies and the necessarily unknown future developments:- a one-size-fits-all approach does not apply- ethical guidelines will have to be regularly re-debated and updated

AI Technologies and Services, which would be in charge of implementing this societal control, including post hoc return on incidents for individual users. Instead of weakening it, the results of increased civic engagement in the development of AI would help fostering the trust of society in AI technologies by deconstructing certain myths and providing substance to many of the complex issues outlined in the HLEG AI Working document. The proposal to involve people belonging to minorities and specific demographics to reduce the risk of reinforcing discriminative patterns present in society would be part of such a mechanism.---=== The principle of autonomy and human agency ==The principle of autonomy and human agency are fundamental to the AI debate. The document rightly identifies the need to guarantee the right of people to know whether they are interacting directly or indirectly with AI systems, their right to know and reject being subject to direct or indirect AI decision making and their right to opt out and withdraw. It would however be of utmost importance to complement this aspect with the concept of human supremacy over AI decision-making. Although the idea is expressed to some extent in other chapters, the principles of autonomy and "do no harm" have to fully encompass this idea. When situations become critical, e.g. when lives are in danger, when the risk element comes into question, when non-quantifiable moral dilemmas enter into play, humans have to retain control. It is therefore necessary that regulation and intervention by humans remains possible at all times. ---== The principle of explicability ==The working paper rightly elevates the principle of explicability to one of the key principles upon which to base the development of AI in the future. We welcome putting stress on such an important dimension. However, this section as well considers "informed consent" as a basis for usage of AI services and as experience shows, this is not enough. Furthermore, the document proposes that informed consent be based on the possibility for individuals or groups to request evidence about the instructions and inputs that lead to a certain output, the organisations involved, etc. Instead of considering this an option, proper intelligible explanatory mechanisms on the main parameters, instructions and inputs, their correlation to the outputs, and the role and responsibility of all actors involved in the AI decision in question should become the rule. This would also ensure that the outcome serves the user rather than the commercial interests of certain actors, including third parties, at the expense of users. Without the availability of such explanatory mechanisms, traceability will be undermined and responsibilities diluted. However, as discussed in later chapters of the document, research in explainable AI is in its infancy. This is why, once again, societal control is of utmost importance. ---== Long term risks ==Considering the last section of this paragraph, it is highly alarming that the HLEG AI cannot reach a consensus on threats as basic as the ones listed in the text. Many of these threats are well documented and should not spark controversy, but rather to trigger the finding of responsible answers. One has the intuition that many of these controversies are linked to the tremendous economic and strategic

remain at the level of laudable intentions which however, cannot be yet be followed by deeds. This increases the importance of societal control and debate about the level of risk that we, as a society, are willing to accept. Indeed, as long as explainability remains poor, increasingly pressing questions will surface regarding the legitimacy of AI decisions, given their (sometimes very difficult to detect) impact on individuals and on society as a whole. The existing case of discriminative biases clearly demonstrate this.Furthermore, as mentioned in the previous chapter already, human oversight and the possibility for humans to intervene is fundamental. This is all the more true in critical applications where uncertainty and risk or the presence of moral dilemmas that cannot be expressed in terms of quantifiable parameters require a decision based on human judgment. == Technical and non-technical ways to achieve trustworthy AI ==The second part of chapter II concerns technical and non-technical ways to achieve trustworthy AI. Without diving into technical considerations, the lacking third dimension of the definition of "trustworthy AI" throughout the document – societal control - also echoes in this section. Naturally, creating secure architectures with fallback mechanisms, testing of systems and their auditability are important. However, since the explainability and the traceability of AI systems is difficult to guarantee, technical approaches aimed at avoiding issues ex ante have to be complemented with post hoc evaluation of usage by consumers and a systematized integration of their feedback into an ongoing societal oversight and debate related to AI applications. When it comes to the non-technical ways discussed in the working paper, the EHF believes that, in the sector of AI, the importance of safeguarding ethical and democratic principles, it is difficult to see how certain aspects can be guaranteed without regulation. We will follow with interest the second deliverable of the HLEG AI. The content of that deliverable will complement this one and a final opinion on their joint relevance will be possible when both documents are finalized. In any case, responses to the exposed issues – whether these are codes of conduct or standardization –have to be prompt and be carried out at European level so as to make it possible to leverage the weight of the Single Market and impose a set of high ethical standards at global scale.As expressed throughout our response to this consultation, we propose the creation of a European Observatory of the use of AI, including the design and management of feedback systems allowing to flag incidents, in a similar way as it already exists in sectors of high risk technologies (e.g. nuclear, aeronautical).Furthermore, the entire "algorithmic chain", from the algorithm designer to the professional user, including engineers, data scientists or coders must receive training on the ethical dimension of their sector. Such trainings should highlight the need for transparency, traceability and intelligibility of systems. As highlighted by the HLEG AI, programmes aiming at increasing diversity in in design teams would also have a positive contribution.Finally, as recognized in the document, citizens must be aware of the functioning, the problems and the risks

any stage of the design, deployment or use, including post hoc end-user return of incidents.Furthermore, empowerment of citizens via education, awareness raising on the one hand and massive improvement in user interfaces on the other is fundamental. Unidirectional informed consent is not enough if one wants to help citizens truly understand what parameters, data, inputs and processes influence the outcomes of the AI application they are using.The EHF will follow with great interest the development of this document as well as the drafting of the HLEG AI's other key deliverable concerning regulation.

benefits that AI promises to those who manage to establish themselves in the global market. While the economic incentive is understandable, it cannot overshadow the strict requirement to abide by our democratic principles, values and rights as described in the first 4 sections of this chapter. Furthermore, certain threats seems to have been relegated to mere technical issues, to be dealt with in chapter 2, such as the issue of discriminative biases resulting from social data carrying discriminative tendencies present in society. >> Reinforcing discriminative patterns present in society <<From a humanist point of view, one of the main risk concerns the possible reinforcement of forms of discrimination and the possible picking up by algorithms of reactionary and exclusionary social stereotypes.An algorithm may be conceived biased from the beginning, as a conscious or unconscious consequence of bias held by its makers.That was seemingly the case of a facial recognition software introduced by Google where a young African-American couple realised that one of their photos had been tagged under the “gorilla” tag. The explanation lied in the data with which the algorithm was trained to recognize people. In this case, it is likely that it mainly, if not exclusively, consisted of pictures of white people. As a result, the algorithm considered that a black person had more similarity to the “gorilla” object that it had been trained to recognize than to the “human” object.In other cases, it may be unclear whether the bias and discrimination are the result of the algorithm itself or of its interaction with users.That is the case of the gender bias revealed in the functioning of “Adsense”, Google's advertising platform. In 2015, researchers from the Carnegie Mellon University and the International Computer Science Institute highlighted that it was biased at the expense of women. Using a software called “Adfisher”, they created 17,000 profiles and simulated web browsing to conduct a series of experiments. They found out that women were systematically offered lower paid jobs than men with a similar level of qualification and experience. The precise causes are difficult to establish. It is of course conceivable that such a bias was the result of the will of the advertisers themselves: they would then deliberately choose to send different offers to men and women. It is however also possible that this phenomenon is the result of the algorithm’s learning process. In this case, men may on average have been more inclined to click on ads advertising the highest paid jobs, whereas women would have resorted to self-restrain, following mechanisms that are well known and described in social sciences. Therefore, the sexist bias resulting from the functioning of the algorithm would be nothing more than the reproduction of a pre-existing bias in society.In other cases, the discriminatory result may be totally unintentional. In April 2016, it was revealed that Amazon had excluded from one of its new services (free home delivery in 24h) neighbourhoods mainly populated by disadvantaged people in Boston, Atlanta, Chicago, Dallas, New York and Washington. Initially, an algorithm from Amazon had found, by analyzing the data at its disposal, that the neighbourhoods in question offered little opportunity for profit to the company. Even though Amazon's objective was

related to artificial intelligence. This implies that school curricula raise their awareness about the reality of algorithms and promote genuine education in terms of values, citizenship and critical thinking. Beyond school, public authorities must develop awareness programs on these issues and foster public debate on artificial intelligence in general. This should be a priority of EU policies in the domain of AI.

certainly not that of excluding any particular area from its services because of their predominantly black population, this proved to be the result of the use of this algorithm. It is therefore obvious that Amazon's algorithm had the effect of reproducing pre-existing discriminations, even if no intentional racism was here at work.Even more evident of a non-intentional result was the case of Microsoft’s Tay, a “learning” robot supposed to enter into conversations on Twitter. In less than 24 hours, Tay converted from its humanist and politically correct original attitude to a racist, sexist and xenophobic discourse, as a consequence of its interaction with what people were writing in their responses. Microsoft apologized and recalled that Tay had been built on the basis of “cleaned up” and “filtered” public data. This ex ante precaution clearly turned out not to be sufficient, once the algorithm was left to operate “autonomously” on Twitter and in interaction with other non-proprietary data. This poses a real question: how to train algorithms and AI to use public data without incorporating the worst traits of humanity?We should therefore be aware that the risk of AI becoming the vehicle for reinforced bias and discrimination may depend: 1) on the choices made by the programmers that create the algorithm; 2) on the data absorbed by the system in its interaction with the public; 3) on the simple circumstance that sometimes the “logical” choice is inconsistent with our ethical and constitutional values. >> Commercial interests of third parties – the example of medicine <<Even if, in the future, final decisions will be (and should be) taken by people, a technical pre-structuring and influencing of these decisions will be possible, if not even likely. The opportunities to support medical decisions and therapies that AI offers are promising, and sometimes breathtaking. This concerns the future of clinical care as well as care. We can assume that AI systems will bring about relevant changes in this area. However, especially with regard to patients' autonomy, we should be careful and avoid AI recommendations and decisions that are subject to bias.AI should work for the benefit of human beings. Regarding medical ethics, it necessarily implies that the decision on the appropriate therapy must be based on knowledge and analysis and not depend on the potential benefits of third party interests. Given the existence of current unfair business practices, it is reasonable to highlight the danger of cases where therapeutic choices would be influenced by a selective use of data or (hidden) algorithms that would include the economic interests of health insurances or health care institutions.In order to avoid this, the functioning of AI in the medical field must be of the utmost transparency and explainability. This does not only apply to general therapeutic decisions or procedures, but also to situations at the end of life. For it is precisely here that the individual, autonomous and responsible will of the patient must be the decisive criterion. In this particularly vulnerable and complex ethical situation, the patient’s will is to be respected in the widest possible sense. It is critical to guarantee that algorithms cannot hinder or make impossible the implementation of the will of the patient because ideological

					convictions of third parties or economic interests of institutions become decisive, perhaps without this even becoming apparent. >> Other domains <<Further domains raise even more questions and seems to require more in-depth reflection, debate and deliberations. This is the case for instance of Lethal Autonomous Weapon Systems. The ethical challenges related to this field of application are enormous, especially when one considers the extreme economic and strategic benefits involved. The fact that the HLEG AI has not reached consensus on this question is very concerning. More importantly however, the critical nature of the question suggests that it requires a much wider societal debate. Such a debate could be steered by the EU Observatory on AI proposed by the EHF. A similar question is linked to the way terrorist organisations use existing AI algorithms used to track user preferences and tailor ad contents for purposes other than what they were originally designed for.		
Fiona	Willis	AFME	We would like to raise the following items for consideration:Glossary: As with any other section of the Guidelines that is incomplete at this stage, we encourage the HLEG to ensure that changes to the glossary, including any revisions or additional terms are made available for public consultation before being finalised, to avoid the risk of the inclusion of definitions that may not be applicable to the very broad uses of AI. The role of AI ethics: It would be helpful to emphasise in this section the subjectivity of 'ethics' as a concept, varying between individuals and cultures if the HLEG intends to successfully foster reflection and discussion at a global level. With reference to our 'purpose and scope' comments below, this should be taken into account in the form of flexibility for industry regulators and individual firms to apply the Guidelines to their individual situations, using them as part of their decision-making process for the use of AI. Benefits of AI: The list of possible benefits of AI on page 1 refer to specific use cases, rather than the broader possibilities for the technology across all industries. We suggest that the Guidelines should acknowledge that AI has the potential to benefit all aspects of EU citizens and industry sectors, rather than starting from a narrow position, such as the examples listed on transportation, social welfare, climate change and natural resources. Purpose and scope: It is unclear from this consultation exactly what status the final Guidelines will	Overall, we support many of the concepts outlined in this section. AI has the potential to bring many positive impacts for the financial services industry and Europe as a whole, and an ethical approach to AI should maximise the benefits for all. With this in mind, we would like to raise the following items for consideration: Section 1- The EU's Rights-based Approach to AI Ethics:We are concerned that the statement that 'adopting a rights-based approach will limit regulatory uncertainty' could lead to some ambiguity. In our experience regulatory uncertainty is limited by considered and proportionate legislation, created in consultation with the relevant industry, and by ongoing dialogue with relevant regulators. From the perspective of AI in capital markets, use of AI is already covered by a number of existing regulations as part of a wider-framework of technology-agnostic, outcomes-based requirements.Section 2 – From Fundamental Rights to Principles and Values:Defining "ethical purpose" through three discreet, yet interconnected, themes is complex and may not be easily understood by all persons that are required to interpret the Guidelines. We suggest that a simpler definition of ethical purpose should be developed that can be more easily consumed by all persons, including laypersons, that will need to refer to the Guidelines. In this respect, the Guidelines should mirror their own requirements for AI in being "comprehensible and intelligible by human	AFME commends the HLEG in seeking to design practical implementation guidelines for firms using AI. A principles-based approach to emerging technologies is appropriate to balance innovation risk and security. We are largely in agreement with the principles proposed and their intentions. With this in mind, we would like to raise the following items for consideration. Section 1 – Requirements of trustworthy AI:Data governance: While AFME agrees that datasets may contain biases, it should not be assumed that this is 'inevitable' and/or that complete removal of all bias is a prerequisite for the use of such data within an AI application. Instead, such risks should be mitigated, including via ongoing assessment of the application's outputs. Governance of AI autonomy: While human oversight will always remain important, it should not be assumed that "…the greater degree of autonomy that is given to an AI system, the more extensive testing and stricter governance is required…". AFME believes this should be decided by an appropriate assessment framework, dependent on the system and/or industry. For example, greater governance and human oversight may be more appropriate for systems that interact directly with humans, rather than by the AI system's overall level of autonomy. AFME agrees with the statement that 'different levels or instances of governance' (including human oversight) will be necessary.Non-discrimination: We support	AFME agrees that assessment of Trustworthy AI will be important and welcomes the initial template from the HLEG as a draft for further consideration. Assessments of this nature will ensure that the benefits of the technology can be maximised while minimising the risks, and that the ethical considerations expressed in the Guidelines are addressed. AFME believes that as with all new and developing technologies, it is important that the risks are considered and actively managed. A robust control framework, similar to those that are already in place for other technologies, should be a priority for any capital markets institution investing in the many forms of AI.However, AFME believes that at this stage defining detailed assessment criteria in the form of a prescriptive checklist requires further stakeholder engagement. AFME believes that while high-level principles are useful (such as the MAS FEAT Principles), detailed assessment criteria will need to be defined at both an industry and individual firm level, as it relates to the type and use of AI systems. Attempting to create overarching assessment criteria at this stage may inhibit firms adopting AI in the early stage of maturity or leave some important areas not fully considered sufficiently.Accountability: AFME agrees that accountability for AI is integral for its ongoing use. Each firm's framework of governance and risk management should ensure accountability for the establishment of, and decisions	AFME commends the European Commission in appointing the High-Level Expert Group on Artificial Intelligence and establishing a forum – the European AI Alliance – to engage a broad and open discussion on the strategic importance of AI in Europe and globally. AFME welcomes the first step of the HLEG to draft AI guidelines on ethics as communicated in the March 2018 European Initiative on AI . This a complex and challenging topic which requires significant discussion and input from a wide range of participants. AFME looks forward to engaging further with the HLEG's final Guidelines and its upcoming work on Policy and Investment Recommendations.As with many industries, the application of AI has the potential to transform capital markets and is already impacting many aspects of how the industry operates, from trading and client interactions to risk management and operational processing. However, AI is a rapidly evolving technology that could have far reaching impacts on society. Care must be taken to ensure its use conforms to appropriate ethical standards applied within individual banks and does not unintentionally harm the market or clients. Equally, policy or regulatory frameworks must be supportive of the development of AI as to not stifle innovation and the potential benefits, while maintaining the appropriate balance against market and consumer protection.Capital markets banks have existing codes of business conduct which

have. Under 'Scope of the Guidelines', it is noted that they should not be a substitute to any form of policy-making, regulation, or internal guidelines, and are not an official European Commission Document or legally binding. However, section B on page 3 status that '…it is important that AI developers, deployers and users also take actions and responsibility to actually implement these principles…". Further clarity would be welcomed that these Guidelines will be voluntary, as well as on the nature and timing of the attestation mechanism that will be used. As with the Glossary above, we are concerned that there will be no opportunity to input into the design of the mechanism, to ensure that it works across a broad range of AI users.

beings at varying levels of comprehension and expertise".Section 3 – Fundamental Rights of Human Beings: Equality, non-discrimination and solidarity/the principle of justice: While fairness remains a key measure, it is important to note that fairness should not necessarily mean equality, i.e. that the AI application delivers the same output for all individuals or groups. This would impact on the effectiveness of AI models; whose results respond to mathematic processes on the input data.Section 4 – Ethical Principles in the Context of AI and Correlating Values: The principle of non-maleficence (1): While AFME agrees that the aim of AI should be to 'do no harm', the definition of harm should be carefully considered. We suggest that it should instead be amended to 'prevent harm'. For instance, if a firm uses an AI application to perform suitability checks on its clients, it should not be considered harmful to withhold services from clients that do not pass the assessment. Indeed, the ramifications of providing unsuitable services to individuals or groups can be significant and harmful to society more broadly. Furthermore, we consider that the principles of justice and explicability could be subsumed under this principle. The principle of non-maleficence (2): We are concerned by the use of the term 'negative profiling'. The activity of profiling is not in itself sinister. However, care should be taken that the processing of data on an individual, which may include profiling, does not have a negative impact on the individual. This obligation is in accordance with Article 22 of the General Data Protection Regulation (GDPR – Regulation 2016/679). The principle of autonomy (1): There are situations in which it may be extremely important for an individual interacting with an AI application to be subordinate to that application (while human oversight of the application as a whole is maintained). For example, it may be necessary to prevent certain individuals from over-riding AI applications concerned with safety systems or the detection of crime, such as anti-money laundering (AML). The principle of autonomy (2): We agree that it is important, and indeed mandated under GDPR, that individuals should have the right not to be subject to solely automated decision making. However, this right does not preclude an automated recommendation being taken into account when the ultimate decision is made by a human. The principle of autonomy (3): In addition, the extent to which an individual need to be made aware that they may be interacting with an AI application may depend on the function of that application and the materiality of the impact of that knowledge. For example, it may not be necessarily important to an individual to know that they are interacting with a 'chatbot' rather than a human when a firm is providing certain types of customer service. The principle of explicability: While transparency may be crucially important for some applications of AI, the extent to which it is necessary to be able to explain the internal workings or decision logic of an AI application will vary depending on the function that application is performing. For example, an AI application that routes trade exceptions (e.g. a failed trade) to an operational process within a firm may not require a significant degree of transparency,

the statement that while it may be possible to remove bias from data, bias is inherent. It is important to acknowledge that, while mitigation for bias is a key standard for AI development, it is impractical to require the removal of all bias. A good test might be that use of the AI application leads to less bias than an alternative system or human process would.Robustness: While contingency plans are an important part of any technology governance strategy, it should be considered that two of the key benefits of AI are the speed and scale at which data can be processed. It may therefore be that, in the event of a systems outage, humans would be unable to partially or fully backfill an AI system. The Guidelines should consider that robustness may be achieved by other means, and not just through human backfill. Transparency: As above, we note that the ability to explain the internal workings or decision logic of an AI application will vary depending on the function that application is performing.Section 2 – Technical and non-technical methods to realise trustworthy AI:AFME agrees with the assessment that both technical and non-technical methods must be used, and that good governance of AI should involve a continuous process of assessment and adjustment. Traceability and auditability: We are concerned by the statement that "laypersons should be able to understand the causality of the algorithmic decision-making process and how it is implemented by organisations that deploy the AI system". AI has the potential to deliver huge benefits in a wide range of applications, but in some cases may be a complex technology. While proximate explanations (for individual decisions) are sometimes possible, global explanations of the algorithm, especially if in non-symbolic language for laypersons, are often not. As noted above, the explainability of AI will vary depending on the use to which it is being put. Firms should instead focus on developing sufficient understanding of the technology at management level and within control functions, to ensure appropriate oversight and governance.Regulation: It is crucially important that regulatory bodies develop the skills and resources to respond to and support the development of AI within their industries. This will also allow development of AI as a regulatory tool, for example for assessing large quantities of data or predicting the build-up of risk.Standardisation: AFME agrees that greater standardisation of terms and frameworks related to AI would be of great benefit. Given the cross-border nature of many industries and firms, it would be most useful if such standardisation occurred at a global level, considering initiatives taking place in other jurisdictions. Codes of conduct: As noted above, there is a lack of clarity as to the exact status of this document and how the adherence process is intended to work in practice. Further consultation on this would be welcome. Education and awareness: This is already a key priority for the capital markets industry. As the possible applications and benefits of AI expand, capital markets banks are increasingly investing in AI education and training for staff across their businesses.

involved in, each use of AI and for setting principles for implementation of policies, procedures and the allocation of responsibilities. For example, the assessment list has items related to governance (as it relates to human oversight, responsibility and accountability) in multiple requirements - Accountability, Data Governance, Governing AI Autonomy. AFME suggests that all governance related considerations should be consolidated under one requirement for consistency. Respect for privacy: We are concerned that the questions listed under this section may not be specific enough for an AI contact. Consideration of privacy concerns should go beyond compliance with GDPR or issues of consent. Respect for human autonomy: AFME suggest that this section more closely relates to the requirements for Transparency and for simplicity should be considered under that header. In addition, bullet four refers to users of AI having the facility to 'interrogate' algorithmic decisions in order to fully understand their purpose and data used. AFME suggests that it would be more appropriate for the AI system owner to be responsible for providing, on request, clear explanations and information related to an AI decision that relates to a user.Robustness: AFME suggests that, as with all current technologies, the forms of attack that may impact an AI system will be broad, and may be both internal (for example, an insider threat within an organisation where the AI system resides), or external (for example, a cyber-hack). While many attack scenarios can be mitigated, we believe it is important to emphasise that it is a continuous process for firms to remain resilient within a dynamic threat landscape.

include ethical principles or have separate, dedicated codes of ethics. These codes outline the responsibilities and obligations on a bank's individual employees' and on the overall bank, covering areas such as: complying with applicable laws and regulations; exercising fair judgement; and executing activities openly and fairly. They are designed to address significant risks that banks face, such as systemic, customer and reputational risks, and are reviewed regularly to ensure that they keep pace with developments in technology and markets and with shifts in ethical and cultural expectations. The AFME responses to the sections posed by the Consultation are outlined below. Overall, we feel that the structure and content of the document may require further refinement in order to more clearly identify, and simplify, key concepts and recommendations. We believe that the Guidelines should more readily apply to the broadest application of relevant industries and AI use cases. We also believe that too quickly prescribing formal requirements and assessment criteria may fail to capture, or limit the maturity and continued adoption of, AI. For instance, it is not the case that AI applications that do not immediately meet the principles outlined in these Guidelines should be prohibited, but that further analysis may be necessary. These Guidelines should remain voluntary and follow the collection of wider stakeholder input at an industry level. Given the tight timeframes for completion of the Guidelines, we encourage the HLEG to consider an additional consultation on Chapters II and III, to ensure that the diverse impacted sectors and interest groups have an opportunity to provide input. We suggest that this part of the paper be adopted in its final form at the same time as the Policy and Investment Recommendations. This consultation could be launched at the same time as the finalisation of Chapter I. If Chapter II and III are to be adopted at the same time as Chapter I we would like the paper to restate that these documents are intended as a living document.Finally, we request confirmation that these Guidelines will be voluntary, and further clarity on the nature and timing of the attestation mechanism that will be used. We encourage the HLEG to ensure an opportunity for the public to review and provide input into the design of this mechanism to ensure that it works across a broad range of AI users.We would be pleased to discuss the content of this response further.

			provided that incorrect outcomes can be amended, and the application can learn from those amendments. There are important use cases where a lack of transparency in the decision making process of an AI provides a level of security, accuracy and fairness, for example applications that detect possible financial crime, cyber incidents or terrorist financing. It should also be borne in mind that AI is a technology which should augment, rather than replace, humans. Given that human decision making is not always transparent or fully explained, it would be better to frame this principle in terms of trust. Section 5 – Critical Concerns Raised by AI:We agree with the assessment that there may be situations in which it is important for AI systems to identify individuals, particularly the examples given of detection of fraud, money-laundering or terrorist financing. As above, we note that consideration should be given to the extent to which an individual needs to be made aware that they may be interacting with an AI application.Identification without consent: We suggest that the Guidelines' wording in relation to GDPR Article 6 should be slightly revised, as the Guidelines currently suggest that data processing is only valid to meet a legal obligation. However, GDPR Article 6 lists several bases for lawful data processing, of which compliance with a legal obligation is only one. We believe that the data processing requirements in GDPR are sufficient, and that it may be more appropriate to include a general statement stating this within the Guidelines. This would also future-proof the Guidelines in the event that amendments are made to the GDPR.				
Anonymous	Anonymous	Anonymous	General comments on Executive Summary: • No mention of manufacturing and industrial AI, which gives 25% of GDP • The AI opportunities, benefit with positive outcomes for our environment (e.g. reduction of emissions, use for circular economy, etc) and European economy (e.g. process improvements in manufacturing, traffic improvement, new business models) following ethical guidelines is missing • Siemens, and many other large industry players, have decades of experiences with AI in EU: AI is not a completely new technology but there are decades of experiences with AI that should be reflected. This might also help to calibrate societal concerns, presenting AI not only as disruptive and transformational but also as an incrementally developing technology. • The statement that "no legal vacuum currently exists, as Europe already has regulation in place that applies to AI" should be highlighted more prominently in the document. And it could be made more concrete referring to the existing regulations and EU directives. • Glossary, part A: "Design for all" or "one-size fits all" concept is not applicable to industrial AI applications; therefore, it should not be taken as a guiding principle. • Glossary, Part B: For these reasons, it must be carefully judged in which cases auditing etc. is really necessary and where not. It should also be clearly recognized that there are many AI applications which are completely harmless from an ethical point of	• On section 2. the paragraph on "Informed consent" is not strong enough. According to this formulation, Facebook and Cambridge Analytica are fully compliant. Therefore, a better formulation would be as in the GDPR text, consented data should be used for the consented purpose only. • Section 4: the selected four overarching principles are very useful, and should get a more prominent role in the document • Delete Section 5.5: It is not concrete enough and may lead to misunderstandings. General concerns about unknown future scenarios should not be part of this paper. In the introduction the differentiation between existing weak AI, focused in this paper, and potential future strong AI could make this explicit.	• "Design for all" or "one-size fits all" concept is not applicable to industrial AI applications; therefore, it should not be taken as a guiding principle. • Respect for privacy and security should be replaced by "Security and Privacy by design and by default" • Section 2, Technical methods: excellently written, crisp, to the point and touching the right points. The author should be consulted and/or invited to review the entire document! However, on Traceability and Auditability, generally, it could be difficult for a lay person to fully understand the causality and decision-making process. • Section 2, Non-technical methods: should add another method/paragraph on: School education, vocational training, required curricula & skills for the new and working generations. • High requirements regarding traceability and explainability for laypersons can limit the design of innovative value-added solutions with AI too much. We should avoid that companies are forced to give access to their algorithms because they can contain business secrets and should be protected as IP. In contrast, innovation-friendly, use-case-specific approaches should be chosen, for example those that focus not on the user (people) but on pure technical processes.	• We have to ensure to position Europe also with respect to AI into a leading position and therefore we have to be careful not to establish regulations which negatively influence our possibilities. Adequate ethical guidelines will support this goal. • We should be careful that also SMEs, the backbone of our economy, will be able to follow and implement these guidelines overall. • Fostering R&I on achieving Trustworthy AI in EU should get a prominent position in the document, at the forefront of activities, including practical test cases (e.g. sandboxes) for various verticals. • The requirements for AI development processes vary depending on the application and the specific solution. In order to avoid complicating or slowing down the development of AI by excessive overhead processes, the procedures should be designed pragmatically and application-specific. Companies should set up these processes themselves in a requirement-oriented manner.	While we fully support the intention of creating trustworthy AI, we would demand guidelines, which are effective and efficient in avoiding the social risks of AI while minimizing as much as possible the overhead on the development and deployment process for AI enabled systems. To implement AI successfully in Europe it must safeguarded that no unbalanced additional costs and bureaucracy is established, e.g. the additional auditing services and storage of logs would demand additional development effort, operational cost for storage, processing power, licenses, as well as man power to monitor and maintain the auditing the system. Especially respecting the possibilities of the huge amount of SMEs in Europe. "Design for all" or "one-size fits all" concept is not applicable to industrial AI applications; therefore, it should not be taken as a guiding principle. From our point of view it is of outmost importance to distinguish between AI applications in the B2C and B2B business and to reflect the degree of human involvement or simple technical needs. The specific recommendations and guidelines should be more aligned to existing processes for security, data protection, product safety and security. Aligning it would mean to make a fit/gap analysis with existing procedures and integrate it into them in a lean manner (e.g. adding further constraints to existing process rather than creating completely new one, wherever possible). • We would appreciate the document as a first impulse for future discussions.

			view, e.g. a passive anomaly monitoring system, that analyzes vibrations of a motor or pump and gives an alert if an anomaly is detected. The majority of application at Siemens, or other large industrial players, is of this kind. Any guidelines should explicitly acknowledge that such applications need not be monitored e.g. by an ethics committee. Then the ethics efforts can be concentrated on those which really pose severe ethical problems.				
Liselotte	ENGSTAM	4Boards.ai Project led by Chalmers with FCG, Combient, FCG, Innovisa, Digoshen, IMIT	Current text - Trust in AI includes: trust in the technology, through the way it is built and used by humans beings; trust in the rules, laws and norms that govern AI – it should be noted that no legal vacuum currently exists, as Europe already has regulation in place that applies to AI – or trust in the business and public governance models of AI services, products and manufacturers. - Suggested change: from Business Governance to Corporate Governance (as an established field with self-regulated mechanisms in European Countries - Corporate Governance Codes)	None specific	Accountability Governance Currently - Organisations should set up an internal governance framework to ensure accountability. This can, for instance, include the appointment of a person in charge of ethics issues as they relate to AI, an internal ethics panel or board, and/or an external ethics panel or board. Amongst the possible roles of such a person, panel or board, is to provide oversight on issues that may arise and provide advice throughout the process. This can be in addition to, but cannot replace, legal and compliance oversight; for example, in the form of a data protection officer or equivalent. - Suggest addition - AI Governance Frameworks needs to be included in and aligned to companies Corporate Governance process. This to ensure the appropriateness of the governance framework and see to that it provides an adequate support for an effective and robust decision-making process when it comes to implementation of AI systems. Organizations should also consider establishing that the board, or a board committee such as the ethical board, shall approve the use of AI systems that are introduced in any process, product or service that is essential for the organization or which in any other way is considered to constitute a potentially high ethical risk for the organization. - Suggest addition - Current Corporate Governance regimes are highly self-regulated in European Countries, with separate bodies publishing Country specific Corporate Governance Codes, which companies are obliged to respond "Comply or Explain" to. It is recommended that Country Corporate Governance Bodies revisit their country specific corporate governance codes to include reference to AI governance and accountability. Under Key guidelines Currently - Ensure a specific process for accountability governance. Suggest addition - Ensure a specific process for accountability governance, aligned with corporate governance where applicable.	Under Assessment list Accountability * Ensure that AI governing process is included and aligned within the Corporate Governance process * Ensure accountability owned at highest level, for companies at the board level * Ensure education of responsibility, and that transparency of AI principles and usage is shared with highest responsibility (for businesses the corporate boards) * Ensure boards consideration of updating corporate policies with AI principles General Comments In general our provided comments are around the corporate boards responsibility for AI governance, which is little referred to in the document. *boards as final responsible for corporations should be educated in the field of AI and the ethical aspects of the use of these techniques and have a way to guide their organisations, also to include it in potential policies and code of conducts. *and that countries' corporate governance codes should include references to AI governance in order to integrate AI governance as a part of the corporate governance, as major part of corporate governance is not regulated but governed under country specific Corporate Governance Codes with "comply or explain" obligation by companies. It is somewhat unclear what the status of the guidelines are, as they are not intended to be binding. It is mentioned that the intention is that they can be "endorsed" by organisations and companies, but it is not further described how such endorsement will work in practise. The guidelines would benefit from this being further elaborated. As there are many organisations publishing different version of AI ethical guidelines it would be helpful with some comparison across these, similar to different sustainability reporting guidelines.	Input on Definition of AI document; • The definition introduces three "flavors" of Machine Learning, but only two (Supervised and Reinforcement) out of three (Unsupervised Learning) are described (p.4) • Deep learning (p.5) o It's stated that "This (deep learning) makes the overall approach more accurate and with less need of human guidance". A more correct statement would be "This (Deep learning) is a good tool for handling complex data relationships, but does not guarantee increased accuracy in result nor reduced human guidance". o Several layers in a neural network will not automatically give you a better result, rather increase the risk of over-modelling data. • The section on Sensors is vague in its description (p.2). The statements make it seem the only thing you need to do is to send pictures to the system, and it will take care of the rest. However, in reality, one must first understand what format is required, then translate the picture into the required format (e.g. pixels) before submitting the picture. • On P.5 there is a statement that reads "(that is, to minimize the error between the expected output and the output computed by the network)". This is the same thing. Instead, it should say "(that is, to minimize the error between the actual outcome and the output computed by the network)". Some information about us sharing this feedback; We are currently performing a 2-year research project 4Boards.ai - led by Chalmers University of Technology in Sweden, together with organisations Combient, FCG, Innovisa, Digoshen and IMIT. The research project 4Boards.ai is exploring best practices on how to enable corporate boards to more successfully govern and leverage AI and other exponential technologies in their innovation and sustainability efforts.
Iain Boyd	Richmond	Deutsche Telekom AG	Europe's digital ethics is a key competitive advantage when compared to other global markets, and the prerequisite for the success of AI is the trust of society in its products & services. Transparency, and hence clearly defined areas of operation for AI must be set by existing laws. Promoting the development of AI, its application, knowledge & social acceptance and driving AI into a future that puts people & hence society first.	The potential of AI can only be realized if it is human centric i.e., a clear focus on ethics throughout its life cycle from R&D to the final application of AI. Private autonomy is the basis of digital sovereignty of every individual, & equally the digital sovereignty of Europe as a common economic market and 'lebensraum,. As stated, many companies are new to the challenges of AI, it is therefore imperative that an energetic social debate on AI is conducted, in particular from the EU across all of its	Realising the trust of AI is of course multifaceted. Under the triad of data protection, usage & access, it is the protection & usage of personal data that is one of the main concerns of individuals. This must be transparent & subject to approved security measures (GDPR provides a sound legal basis for this), and users should retain control of the use of their data. In addition to the protection of personal data, AI systems also raises questions about the transparency of the used algorithms. Human		The societal impact of AI is difficult to quantify. AI & hence digitalization are societal changes that cannot, and should not be stopped, but as a society we must embrace these changes. The right framework for education & employment are, alongside digital ethics, fundamental for the sovereign application of AI. The discussions surrounding the impact of AI, and the socio-economic changes that are already occurring, in particular the changing labor market, should not be defined by contrasting

				member states. This will forge trust in AI, in addressing key AI issues such as opportunities & risks for the economy, and in setting a clear holistic AI ethical framework that all must adhere to. To avoid dissuading AI innovation, where possible, existing legal & regulatory frameworks should be applied i.e., further developed where necessary, although this could in some cases lead to legal uncertainty. A balance between new legislation that does not prevent innovation, but also enhances legal certainty is required.	interaction with AI systems also requires equal transparency, and users should know when they are interacting and/or communicating with an AI system. The unpredictability of AI learning systems requires intense debate as to responsibility/accountability regarding liability and security. Security by design must also apply in the development/application of AI. Communication of security vulnerabilities, security updates & features such as an emergency override function for AI systems all enhance the attribution of clear responsibilities. Clear responsibilities would also augment the issue of liability, which would mitigate lengthy & expensive judicial processes for damages/misuse, but at the same time enhancing legal certainty. Mandatory risk assessments for critical AI applications such as in the health care sector could contribute towards ensuring trust. IT security & product safety of AI applications goes hand in hand, and this correlation must be considered by developers and industrial users alike. Obligatory certification for the application of AI in critical infrastructures and critical IoT solutions could build on the European Certification Framework & follow the CE standard. Finally, the human administrator should always have the possibility to override an AI decision & take immediate corrective action, which is immensely important where life and limb are at risk due to the application of autonomous AI systems.	human & machines, but should rather focus on how humans & machines can work in symbiosis towards the same goals. An AI system can support an employee in his/her role by for e.g., conducting repetitive tasks, leaving the employee free to concentrate on essential tasks, which in turn increases productivity. AI presents opportunities in the changing labor market by increasing for e.g., efficiency, new business models, undertaking mundane/dangerous tasks, which could all lead to new employment opportunities. These changes require constant observation however, so that we can actively shape and assess the changes as quickly as possible. Encouraging digital skills in the education system i.e., schools/universities, with a focus on STEM faculties, IT security, programming etc. are essential for future generations/societies being prepared for the enormous changes that will shape their digital future worlds, but also a focus on ethics in education is paramount & should accompany AI/ICT courses. This is imperative for Europe to be able to compete with Asian and US competitors, who incidentally are investing far more money in the development of AI than Europe.
Anonymous	Anonymous	Anonymous	Glossary - definition of Bias (page iv) - it should be acknowledged that bias is not always a bad thing and is sometimes also intended. Objective should be to make sure there is no unfair discrimination. The role of AI Ethics (page 2) - the document rightfully states it should not be regarded as an end point, but rather as the beginning of a process of discussion. But it should be specified how this process will take place and if this means a continuation of the HLEG.	•Freedom of the individual (page 7) - How does this fit with national security obligations? •Ethical principles of AI (page 8) - what happens if an AI developer is faced with a contradiction between different principles? Is there a hierarchy that needs to be applied? Some guidance is needed. •Principle of "Preserve human agency (page 9) - how realistic is that people should always have a right to opt out and a right of withdrawal? With the increasing use of AI this becomes unrealistic and difficult to implement. Instead this right should be based on the type of AI system (the sensitivity of the use case). •Principle "Operate Transparently" (page 10) - need to make the point that first requirement should be the need to inform individuals on whether or not they are interacting with an AI system, that is the basis of transparency. •Principle "Operate Transparently" (page 10) and "identification without consent" (page 11) - informed consent is important, but under GDPR it is not the only legal basis for data processing...legitimate interest is also allowed.	The chapter on "Realizing Trustworthy AI" contains too many requirements (page 14) and will make it challenging for developers to operationalize them. Therefore we recommend merging some: "data governance" and "respect for privacy" "design for all" and "non-discrimination" "governance of AI Autonomy" and "respect for human autonomy" "robustness" and "safety" Some suggested changes: Data Governance - need to recognize that in certain cases bias is intended because of the objective of the AI system Design for All - it is not realistic to request every AI system to be designed in such a way that it can be used by all Robustness - Need more guidance on what level of accuracy is required for AI systems, especially sensitive use cases. Will the Commission set up a process to develop guidance? Robustness - the fall back plan should depend on the use case, in many use cases this might not be necessary.	The chapter on "Assessing Trustworthy AI" which contains a long list of questions requires more work, need to be clearer and more detailed so they can actually be used by developers. Obviously it does depends on each use case, but besides questions there is also a need to provide suggested answers.	•We welcome the objective of the EU HLEG to develop Ethics Guidelines which are not just a compilation of values and principles to be respected, but which most importantly provide guidance on how to actually implement these in the development and use of AI systems. This will be the real added value of the Ethics Guidelines and the current draft, which is open for consultation, is a very good basis and now needs to be further "operationalized". •Support that "Trustworthy AI" has two components: an ethical purpose and be technically robust. This will make sure that AI is trusted by its users and that it will result in improving Europe's competitiveness. •Welcome the intention to use these Guidelines to foster a reflection and discussion on a global level, the mid-long term objective should be to develop international frameworks. •Welcome the recognition that the implementation of the high level guidelines depends on the use case and therefore a tailored approach is needed, we cannot expect to implement the guidelines in the same way in each use case. •Support the 5 ethical principles that are suggested: do good, do no harm, preserve human agency, be fair, operate transparently •Largely support the content of the 10 requirements for trustworthy AI, but suggest merging some and also making some slight adjustments. •The Technical and Non-Technical Methods listed to achieve Trustworthy AI are a good non-exhaustive list, but more emphasis should be placed on the role of standardization and codes of conduct.

Sam	Work	Best Practice AI								Healthcare Diagnose and Treatment1. Accountability: - Who is accountable if things go wrong? Comments: The definition of "wrong" first must be decided upon, which involves identifying who is responsible for providing that definition. - Are the skills and knowledge present in order to take on the responsibility? (Responsible AI training? Ethical oath?) Comments: Included in this should be training on relevant legal decisions around accountability, particularly in the realm of healthcare, and ensuring an awareness of any relevant legal decisions surrounding the use of AI in healthcare. - Can third parties or employees report potential vulnerabilities, risks or biases, and what processes are in place to handle these issues and reports? Do they have a single contact point to turn to? - Is an (external) auditing of the AI system foreseen? - Was a diversity and inclusiveness policy considered in relation to recruitment and retention of staff working on AI to ensure diversity of background? - Has an Ethical AI review board been established? A mechanism to discuss grey areas? An internal or external panel of experts? Comments: Specifically to this use case, should this fall under the auspices of existing healthcare ethics review boards or should an AI-specific committee or sub-committee be set up? 2. Data governance: - Is proper governance of data and process ensured? What process and procedures were followed to ensure proper data governance? Comments: We recommend re-wording this point to be more instructive: "What process and procedures should be followed to ensure proper data governance?" With regards to healthcare, this includes following pre-existing data collection and processing regulations, with particular sensitivity to where anonymisation can and should be done. - Is an oversight mechanism put in place? Who is ultimately responsible? Comments: We recommend re-wording this point: "What oversight process is in place to ensure adherence to the data governance process and procedures, including the responsible parties for implementing it?"What data governance regulation and legislation are applicable to the AI system? This point should inform the first point, so perhaps should be re-ordered. 3. Design for all: - Is the system equitable in use? Comments: In a healthcare setting it is essential to identify who the users will be and who could potentially need access to the system. - Does the system accommodate a wide range of individual preferences and abilities? - Is the system usable by those with special needs or disabilities, and how was this designed into the system and how is it verified? - What definition(s) of fairness is (are) applicable in the context of the system being developed and/or deployed? - For each measure of fairness applicable, how is it measured and assured? 4. Governing AI autonomy: - Is a process foreseen to allow human control, if needed, in each stage?Comments: For healthcare, this will include policy on when human (e.g. technicians, doctors, etc.) review of machine decisions/recommendations is necessary, and, in the result of a disagreement between the two, what is the policy for whose decision is deferred to? - Is a "stop button" foreseen in case of self-learning AI approaches? In case of prescriptive (autonomous decision making) AI

approaches? - In what ways might the AI system be regarded as autonomous in the sense that it does not rely on human oversight or control? - What measures have been taken to ensure that an AI system always makes decisions that are under the overall responsibility of human beings? Comments: This point is very much a re-statement of the first under “Accountability” and the first point in this section. - What measures are taken to audit and remedy issues related to governing AI autonomy? - Within the organisation who is responsible for verifying that AI systems can and will be used in a manner in which they are properly governed and under the ultimate responsibility of human beings? Comments: Again, this is covered under “Accountability”. 5. Non-discrimination: - What are the sources of decision variability that occur in same execution conditions? Does such variability affect fundamental rights or ethical principals? How is it measured? - Is there a clear basis for trade-offs between conflicting forms of discrimination, if relevant? - Is a strategy in place to avoid creating or reinforcing bias in data and in algorithms? Comments: With regards to healthcare, the training and validation data used should be identified to determine if it is applicable for all patients or if there was insufficient variability in the data in order for the system to provide “accurate” diagnoses or prescribe treatment for all patient groups. An example would be a machine learning system trained to identify imminent risk of heart attack on a data set which only included white males in a certain age group. This is a well-documented problem for medicine in general, but can be exacerbated by an AI system. - Are processes in place to continuously test for such biases during development and usage of the system? - Is it clear, and is it clearly communicated, to whom or to what group issues related to discrimination can be raised, especially when these are raised by users of, or others affected by, the AI system? 6. Respect for Privacy: - If applicable, is the system GDPR compliant? - Is the personal data information flow in the system under control and compliant with existing privacy protection laws? - How can users seek information about valid consent and how can such consent be revoked? - Is it clear, and is it clearly communicated, to whom or to what group issues related to privacy violation can be raised, especially when these are raised by users of, or others affected by, the AI system? 7. Respect for (& Enhancement of) Human Autonomy: - Is the user informed in case of risks on human mental integrity (nudging) by the product? - Is useful and necessary information provided to the user of the service/product to enable the latter to take a decision in full self-determination? - Does the AI system indicate to users that a decision, content, advice, or outcome, is the result of an algorithmic decision of any kind? - Do users have the facility to interrogate algorithmic decisions in order to fully understand their purpose, provenance, the data relied on, etc? 8. Robustness:Comments: Much of this section is more technical answers to how to achieve the previous sections - perhaps there is a way to better organise it to reflect that.Resilience to Attack:What are the forms of attack to which the AI system is vulnerable? Which of these forms of attack

can be mitigated against?What systems are in place to ensure data security and integrity? Comments: Given the nature of healthcare and sensitivity of patient data, this is particularly important with regards to the storage and collection of personal data.Reliability & Reproducibility:Is a strategy in place to monitor and test that my products or services meet goals, purposes and intended applications?Are the used algorithms tested with regards to their reproducibility? Are reproducibility conditions under control? In which specific and sensitive contexts is it necessary to use a different approach?For each aspect of reliability and reproducibility that should be considered, how is it measured and assured? Comments: What is the threshold for inaccurate diagnoses or prescribed courses of treatment? What standard are the human healthcare providers held to, and should the system be expected to meet or exceed the same standard? How should this be measured - for instance, in the case of diagnosing disease, should the goal be to minimise the false positives or minimise the false negatives?Are processes for the testing and verification of the reliability of AI systems clearly documented and operationalised to those tasked with developing and testing an AI system? Comments: Machine learning systems require being updated with new data intermittently. There should be a process put in place for identifying the frequency with which this should happen (e.g. reliance on external vendor's recommendations? Internal best practices?).What mechanisms can be used to assure users of the reliability of an AI system?Accuracy through data usage and control:What definition(s) of accuracy is (are) applicable in the context of the system being developed and/or deployed? For each form of accuracy to be considered how is it measured and assured?Is the data comprehensive enough to complete the task in hand? Is the most recent data used (not out-dated)?What other data sources / models can be added to increase accuracy?What other data sources / models can be used to eliminate bias?What strategy was put in place to measure inclusiveness of the data? Is the data representative enough of the case to be solved?Fall-back plan:What would be the impact of the AI system failing by: Providing wrong results? Being unavailable? Providing societally unacceptable results (e.g. bias)? Comments: For healthcare, planning for unavailable electronic systems and digital data is critical. "Societally unacceptable results" and "bias" must also be defined in the context of healthcare: for example, does the system recommend different courses of treatment for different demographics which may be interpreted as discriminatory by the public?In case of unacceptable impact - Have thresholds and governance for the above scenarios been defined to trigger alternative/fall-back plans?Have fall-back plans been defined and tested? Method of building the algorithmic system- In case of a rule-based AI system, the method of programming the AI system should be clarified (i.e. how they build their model)- In case of a learning-based AI system, the method of training the algorithm should be clarified. This requires information on the data used for this purpose, including: how the data used was gathered; how the data

used was selected (for example if any inclusion or exclusion criteria applied); and was personal data used as an input to train the algorithm? Please specify what types of personal data were used.Method of testing the algorithmic system- In case of a rule-based AI system, the scenario-selection or test cases used in order to test and validate their system should be provided- In case of a learning based model, information about the data used to test the system should be provided, including: how the data used was gathered; how the data used was selected; and was personal data used as an input to train the algorithm? Please specify what types of personal data were used. With regards to externally created systems, was proper data collection adhered to by the creators? Comments: An example of where this has been unclear and resulted in loss of public trust is the UK's NHS data sharing scheme with the company DeepMind to create healthcare solutions (although this may be as much due to UK newspaper editorial views on Google - and its affiliates - as specific issues with that instance of research). Outcomes of the algorithmic system- The outcome(s) of or decision(s) taken by the algorithm should be provided, as well as potential other decisions that would result from different cases (e.g. for other subgroups).

Kalina	BOZHKOVA	MedTech Europe	MedTech Europe welcomes the vision set out by the European Commission to support ethical, secure and cutting-edge AI made in Europe. AI has the potential to substantially improve the delivery of healthcare and other services that advance well-being, provided that it is embraced, supported and trusted by patients, healthcare professionals and other stakeholders. Embedding AI in an ethical framework that respects fundamental rights, principles and values is critical to building this support and trust. Therefore, the two-fold approach of the AI Ethics Guidelines for Trustworthy AI looking at 1) the ethical purpose of AI and at 2) the AI technical robustness and reliability, is appropriate. The AI Ethics Guidelines by the AI High Level Expert Group (AI HLEG) offers robust guidance for businesses, including for the medical technology industry, to develop and implement trustworthy AI solutions. Hence, the proposed mechanism to enable stakeholders to voluntarily endorse and sign up to the Guidelines presents an opportunity to set equal standards across organisations. It should be kept in mind, however, that while the guidelines could be a state-of-the-art mechanism, it might be operationally more challenging to fully integrate those guidelines and go beyond existing fundamental rights legislations.MedTech Europe is also looking forward to the opportunity to comment on the second deliverable of the AI HLEG on Policy & Investment recommendations for AI.	Sections 1,2,3,4MedTech Europe (MTE) shares the AI HLEG's belief in an ethical approach towards AI, based on existing legal instruments like the EU Treaties, Charter of Fundamental Rights and specific acts such as the General Data Protection Regulation (GDPR). MTE also commends the AI HLEG for working on assessing existing ethical principles from different initiatives in the area of technological development and for having identified specific principles and values for AI (i.e. "Do Good", "Do no Harm", "Preserve Human Agency", "Be Fair", "Operate Transparently".). These principles are well known in the healthcare community through existing conventions, from the Nuremberg Code to The Common Rule.Section 5MedTech Europe agrees that the identified potential areas of concern are also valid for the healthcare sector.	Section 1MedTech Europe supports the requirements proposed by the AI HLEG for Trustworthy AI and recommends that "Cybersecurity" be added to the list of ten requirements of trustworthy AI:1) Accountability2) Data Governance3) Design for all4) Governance of AI Autonomy (Human oversight)5) Non-Discrimination6) Respect for (& Enhancement of) Human Autonomy7) Respect for Privacy8) Robustness9) Safety 10) TransparencyData governanceWith reference to the requirements related to Data governance, there are some additional ethical considerations that should be taken into account. Completeness and robustness of data is necessary to avoid ethical issues related to bias and discrimination. Therefore, more efforts should be put in building a health data ecosystem which enables adequate representation from population and therapeutic areas, enhances the transparency and harmonization of data collection, and promotes access and sharing of health data (e.g. data donation) in compliance with the GDPR, for healthcare research and treatment purposes. There are very interesting examples of data sharing that help balance privacy rights with the potential healthcare benefits residing in data, such as:- The research project between the Digital Ethics Lab of the Oxford Internet institute, the Data Ethics Group at The Alan Turing Institute, and Microsoft, exploring the "Ethics of Medical Data and Advanced Analytics" which investigated existing ethical frameworks, such as the "Ethical Code for Posthumous Medical Data Donation".- The Yale University Open Data Access (YODA) Project which aims for the responsible sharing of clinical research data, open science, and research transparency. The YODA Project seeks mutually beneficial partnerships with Data Holders, promoting independence, responsible conduct of research, good stewardship of data, and the generation of knowledge in the best interest of society. To participate, each Data Holder must transfer full jurisdiction over data access to the YODA Project.SafetyWith reference to the requirements related to Safety, it is important to highlight that this is fundamental for medical devices and needs to be differentiated from the security aspects of a medical device. While safety is about ensuring that the system will indeed do what it is supposed to do, without harming users (human physical integrity), resources or the environment, the security is about the technical aspects of an algorithm, which are ensuring it is programmed to handle successfully malicious software. TransparencyTo be acceptable or legitimate, the decisions of an algorithm need to be understood, thus explained. That is why "Transparency" (i.e. "explainability" of how the algorithm delivers its conclusion) is one of the key principles for application of AI in the healthcare sector. CybersecurityA proposed explanation of the role of "Cybersecurity" in Trustworthy AI in the healthcare sector, could be as follows: Cybersecure AI needs human oversight and interpretation, in order to be able to recognize successfully unwanted and unpredicted commands. Human intervention would facilitate the detection of malicious commands when alerted by the algorithm, when automated processes have been insufficient. Cybersecurity plays a vital role	The following use cases of AI in healthcare are examples of how AI is implemented in the healthcare sector in compliance with existing regulations (e.g. In-vitro Diagnostic Regulation (IVDR) 746/2017 & Medical Device Regulation (MDR) 745/2017, General Data Protection Regulation/ GDPR).A helpful consideration while building the right approach for the realization of Trustworthy AI, would be that not all requirements might be fully applicable in each scenario, depending on the intended purpose and use of AI, as well as its operating environment,• AI for diagnosis:An example of AI applied in software training is a medical image analysis software which can help physicians make more accurate clinical decisions using computer support. The software would use machine learning models, trained on large datasets, labelled by healthcare experts. After physicians upload patients' scans into the software, it compares them to the ones labelled to contain anomalies indicating lung cancer, heart disease, or other conditions, which enables physicians to make more precise diagnostic decisions.AI and deep learning also enhance diagnostic possibilities in areas like breast and colon cancer detection, pulmonary diseases, brain tumour segmentation, or Alzheimer disease.• AI for disease management: There are mobile applications available on the market which provide assistance in the management of chronic conditions like diabetes. Using AI the app would evaluate how a user's blood sugar levels respond to variables such as food intake, insulin dosing and other daily routines, in other words, it would predict the likelihood of an individual experiencing a low glucose event. Such applications are integrated in continuous glucose monitoring systems, which are designed to give users predictive alerts up to 60 minutes before they experience hyper- or hypoglycaemia.• AI for robot-assisted surgeries: Robotic Surgical Systems (i.e. computer-assisted surgeries) allow surgeons to perform minimally invasive surgeries with the help of robotic arms. The surgeon operates while seated at a console unit, using hand and foot controls, and with a 3D, high-definition view of the surgical field. The system can simulate an open surgical environment without physical trauma of large incisions. Such surgical systems collect large databases from similar operations, which then, thanks to AI, would allow the performance of unassisted AI surgery with a greater than human precision.We invite the AI HLEG to consider these use cases when tailoring the assessment list to the healthcare sector. MedTech Europe would be happy to contribute to such an exercise, and to provide more information on any of the above briefly explained cases.

in the healthcare sector. Connected devices have greatly improved access in remote areas, allowing for faster and more accurate diagnoses, and aiding the management and transfer of medical records and images. AI and machine learning (ML) can be used for analysis of digital imaging from medical devices, whereas the reliability depends on the quality of the training data, especially when data is related to the electronic health records, the treatment of patients ,or diagnostics. As more hospitals connect medical imaging equipment to the internet, the risk of malicious cyberattacks increases exponentially. Section 2Furthermore, the proposed lists, which include existing and to-be-developed technical and non-technical methods to implement the above requirements, comprehensively capture the needs for realizing Trustworthy AI: - Technical: ethics & rule of law by design; architectures for trustworthy AI; testing and validating; traceability & auditability; explicability; - Non-technical: regulation; standardization; accountability; codes of conduct; ethical education; stakeholder and social dialogue; diversity and inclusive design teams;Since it is mentioned that some of the proposed methods might inform the second deliverable of the AI HLEG (with policy & investment recommendations), it might be helpful to point to some existing sectoral instruments, which ensure a sufficient level of conformity with the non-technical methods. Implementing the full assessment requirements list proposed above could be considered a best-in-class-practice, particularly important within the healthcare sector. It should be acknowledged that many requirements, for example those related to data governance/transparency, safety of patients, healthcare professionals and security (cybersecurity) are already comprehensively addressed in sector specific regulations, like the In-vitro Diagnostic Regulation (IVDR) 746/2017 & Medical Device Regulation (MDR) 745/2017, and more specifically within the accompanying Implementing Acts and Guidance. When software is qualified as an In-vitro Diagnostic or as a Medical device, sets of specific requirements must be met to demonstrate safety, clinical performance and cybersecurity. Particularly for smaller companies seeking to document that they meet the criteria for trustworthy AI, it may be helpful referencing these already established requirements. Finally, the AI Ethics guidelines could be technically enhanced and complemented by existing international standards. Standards like those by ISO IEC on trustworthiness, which are currently a subject of development, could offer an opportunity for more systematic operationalization of the guidelines, and could be used as a mean to demonstrate ethical compliance.

Anonymous	Anonymous	Anonymous		4. Ethical Principles in the Context of AI and Correlating Valuses * The principles of Autonomy: "Preserve Human Agency" SKL välkomnar särskilt skrivningarna om transparens och insyn i hur en process för till exempel träning av AI genomförts. Om en sådan dokumenterad process kan redogöra för hur sådant som preparering av data och test för diskriminering gått till ökar möjligheterna för allmänheten att lita på att den ansvariga organisationen gjort vad den kunnat för att uppfylla principen om icke-diskriminering. En förutsättning för reell transparens är att tillvägagångssättet presenteras på ett sätt som är lättillgängligt och begripligt även för personer som inte är tekniskt kompetenta. En sådan insyn skulle kunna skapas genom att en oberoende part granskar lösningar och presenterar resultatet av sin granskning på ett för allmänheten begripligt sätt. Avsnitt 5 - Critical concerns raised by AI Det är rimligt att anta att det i takt med att tekniken blir mer spridd kommer uppstå flera områden där avvägningar runt användning av AI blir aktuella. Därför skulle det kunna bli nödvändigt att ha en permanent gruppering som löpande kan hantera etiska frågor runt AI.		SKL välkomnar att EU kommissionen tar fram en etisk vägledning för pålitlig AI med stöd av experter på hög nivå och ger stöd för det som tas upp. Trots att vägledningen inte blir bindande kommer den kunna utgöra stöd till nationer och organisationer i det egna arbetet med att ta ansvar för utveckling, distribution och tillämpning av AI. Att detta ansvar tas på alla nivåer och inom alla områden kommer att vara avgörande för hur väl möjligheterna inom AI ska kunna tas tillvara utan att de risker som finns med AI leder till en utveckling som skadar individer och samhället i stort. Code of conduct - är rimligt att alla organisationer med öppna och delade data har framöver. Dessa guidlines kan vara en av utgångspunkterna också för ett sådant internt arbete.
Friederike	Ladenburger	COMECE, Commission of the Bishops' Conferences of the European Union	The High-Level Expert Group explains in the introduction the character of these drafted guidelines. They underline that the guidelines will offer guidance to all relevant stakeholders concerning ethical challenges of AI, they will be not legally binding and they do not intend to be a substitute to any form of policy-making or regulation.The authors define that the scope of the guidelines covers AI applications in general – although they are aware that a tailored approach is needed for the different challenges of AI context-specificity.a. We have to criticize the procedure of publishing the draft of the guidelines. The time for contributions to this consultation is much to short. Publishing the paper just before Christmas with a first deadline for the 18.1.2019 and a postponed deadline for the 1.2.2019 for the contributions is not convincing. The purpose of the drafted guidelines is to invite all different stakeholders to share their opinions and to support the final version. The given deadline is predestined to exclude stakeholders which do not have the capacity for an immediate reaction.b. We also have to criticize the undifferentiated use of the term “stakeholder”. The drafted guidelines are mentioning the problem, that different situations raise different challenges. But as a result of this recognition it is not convincing just to mention the necessity of differentiating business-to-consumer or business-to-business or public to citizen in general, but not differentiating the necessity of balancing out the different interests in the wider context of the common good of these different stakeholders. An ethical assessment has to deal with different factors and tools whether an AI developer, a user or the public sector is affected.	We very much welcome the underlining of the human-centric approach of AI in the EU in the drafted guidelines. AI has to serve the common good. AI has to serve the lives of all human beings. It has to be considered that human life not only has a personal dimension but also a community dimension - community in its human, universal dimension.The structure of Chapter I is misleading. The used terminology is not coherent with EU law. The fundamental rights are seen as the bedrock for the formulation of ethical principles.From our point of view the argumentation should start the other way around:A society’s ethics is based on general values (depending of cultural, societal, anthropological and religious convictions). From these convictions more concrete norms for acting are build up (principles) - they are more concrete and are influencing a society, only some of them are implemented into fundamental rights.The different character of fundamental rights and principles always has to be clear:A fundamental right gives a concrete subject the possibility to defend its position against the state or against other actors with legal means.A principle is a given regularity which gives an ethically order for acting. Legal obligations are not in competition with not legally binding ethical obligations: but of course, there is a clear interaction between law and ethics. In Chapter I, part 3 the list of fundamental rights is not correct and coherent with the use of technical terms in the EU treaties.Part 3.3 is mentioning “Respect for democracy, justice and the rule of law” as a fundamental right. This is not corresponding with the legal terminology of the EU treaties. In Art. 2 EUV “Democracy, justice and rule of law” are mentioned as values.In Chapter I, part 3.4 “solidarity” is mentioned as a fundamental	Chapter II, part 1 names a list of requirements. Most of these terms are named in other documents as principles. (See Statement of the EGE on Artificial Intelligence from March 2018).We would recommend to add to this list the term of the “primacy of the human being”. The draft mentions by itself already this expression in application of the Oviedo Convention.By this requirement the human centered approach of the draft could be underlined and it would be clarified that it is a premise for the personal and community dimension of the existence of the human being.	

					right. This is also not coherent with the terminology of the EU law and with other International binding instruments.The use of a coherent terminology is essential to allow these guidelines to become a useful and concrete ethical tool for business, public sector or citizens.We are skeptical concerning the concept of the principle of “explicability” (Chapter I, part 4)In this term the two components of “transparency” and “accountability” are mixed up. We are concerned that the very important aspect of “accountability”, specially under the important aspect that it has to be user – centered, can be neglected. Our concerns are underlined by the fact, that Chapter II is mentioning “accountability” as a requirement, but not differentiates clearly enough between public and private sector. “Accountability” has to be connected with an “understandability” in order to be human centered. Only if the use of AI is understandable for the user , the user knows which level of transparency and accountability he can demand. Chapter I , Part 5.4 LAWSWith regarding to a possible development and use of artificial intelligence technology in the security & defence domain, it is to be noted that respective EU funding instruments (notably the European Defence Fund) should fully comply with international legal obligations of both the EU and its Member States. Technologies and weapons that are not compatible with the legal standards of international human rights law, international humanitarian law as well as of arms control, disarmament and non-proliferation provisions must not be supported under EU funding.Reaching an international ban on fully autonomous weapons currently seems to be out of scope despite various calls (https://bit.ly/2QCdjVE) and warnings by the scientific community (https://bit.ly/2syu4Hd). Nevertheless, in line with a recent resolution of the European Parliament of 12 September 2018 (https://bitly.com/), the development of ethically problematic technologies, including lethal autonomous weapon systems, should be excluded from EU defence funding. Increase of technological sophistication of weapons tends to disproportionately affect the civilian population (https://bit.ly/2FlFeHA). Fully autonomous weapons enabling lethal actions without meaningful human control pose major legal as well as security concerns. Moreover, the de-humanisation and de-responsibilisation in performing lethal actions raises grave ethical questions (for a deeper reflection on ethical implications of lethal autonomous weapon systems cf. Caritas in Veritate Foundation: “The Humanization of Robots and the Robotization of the Human Person”, https://bit.ly/2QD7UgW).		
Jussi	Mäkinen	Technology Industries of Finland	Technology Industries of Finland (TIF) considers Ethics Guidelines as a fit instrument to pave the way for human-centric application of data and AI to unleash the potential of data-economy and build trust on digital solutions. In order to enhance its effectiveness, the document needs to be significantly shorter and the actionable content needs to be in focus.	TIF points out that many of the aspects covered in this section are already covered by the GDPR. TIF suggests to remove ‘Critical concerns raised by AI’ entirely. It is valuable to underline that humans need to be able to know when they are interacting with AI.	TIF underlines the importance of the quality of data. Quality of the AI systems, solutions and their outcomes is very much affected by the quality of the data. Thus, the quality of the datasets and knowledge on analysis of bias and other data-related issues in of paramount value on this European project. When addressing transparency it is essential to focus on general intelligibility of the outcomes of AI applications, not transparency of neural networks or	TIF regards this operational part as the most valuable one of the paper as it serves as a practical tool to guide developers and deployers to human-centric use of AI. What needs more attention is to better link this document to the GDPR. It does not suffice to refer to compliance with the regulation, but more in-depth analysis is needed. Usage of data will multiply in future and all parties gathering data need to take into account future needs. Europe and European companies need to quickly develop methods	

			Main messages • Within the EU, fundamental choices have already been made to value highly privacy and other fundamental rights of human beings. • What EU needs to lead the way on data-economy is to develop actionable best practices how to combine legal requirements to sound business and technology practices. • There is no distinct set of ethical principles for AI but a need to find a way to apply AI on a way that is compatible with European values. • AI is a domain where there are lots of ungrounded fears and doubts. We need real-life scenarios in order to have fruitful discussion. Human will be in charge of AI systems in foreseeable future. • AI and sustainable use of data have the potential to transform the European societies and businesses to serve humans better than the existing ones.		technological solutions themselves. Basically, what matters is information concerning datasets or categories of data used – not technical details.	to collect and process data on a way that builds trust among people and companies and gives legal clarity on how to utilise data in AI solutions.	
RAMONA	PINEROS	CCOO-FSC afiliated a UNI	- UNI Europa ICTS welcomes the possibility to contribute to the stakeholders' consultation and underlines the importance of a broad public debate and information on AI. This debate must result in clear ethical and social guidelines and standards with the aim of improving the living and working conditions of European citizens.- We acknowledge the innovative potential of AI and new technologies that can be beneficial for our society. However, these new technologies also create challenges and we are concerned about the possible risks and consequences relating to working conditions, skills and training, ethics, equality, health and safety (among others). Therefore, UNI Europa would like to underline the importance of addressing AI technologies and robotization as topics for collective bargaining at all levels (company, national and European). AI and robotics have a huge impact on the future labour market, as jobs will sometimes disappear or be transformed and other jobs will be created. We need to accompany this process and address the question of skills and training for the future workforce: need to ensure that training on necessary digital skills is provided by education institutions and companies, and that it is not the sole responsibility of the worker to keep up with the rapid technological developments. Employability needs to be promoted through upskilling and reskilling schemes for workers. Investment in formal, informal and life-long learning is key; we must enable people to work with AI or invest in competences that AI will not cover. It is important to develop action plans at EU and national level together with education providers and social partners in order to modernize education and vocational training. We therefore welcome the call from the ILO Global Commission on the Future of Work for "a formal recognition of a universal entitlement to lifelong learning and the establishment of an effective lifelong learning system. (https://www.ilo.org/global/topics/future-of-work/publications/WCMS_662410/lang--en/index.htm)- The social partners play a key role in this and the EU should cooperate with them and national governments in order to identify which job sectors will be affected by AI. We need to understand the	- UNI Europa supports the human-centric or human-in-command approach suggested in the guidelines. We agree that it is necessary that humans always need to remain in control of technology and machines. Likewise, we agree that the use of AI needs to respect European values and fundamental rights.- We recommend an expansion of 3.2 "In an AI context, freedom of the individual requires protection from direct or indirect coercion, surveillance, deception or manipulation." Much profiling that lies at the heart of AI systems relies on a degree of manipulation of data. This is not least relevant in relation to the world of work and especially in the use of AI in Human Resources. - We welcome that the HLEG understands the need to ensure that those involved in the development and marketing of AI (researchers, engineers, designers etc.) act in accordance with ethic and social responsibility criteria. This should be addressed by changing educational priorities for technical subjects and by providing lifelong learning opportunities (e.g. by incorporating ethics and the humanities into training courses in engineering). - Organisations and companies should develop tools to facilitate ethical discussions and decision-making throughout the whole design process. This should be completed by internal training programs on ethics for all employees. Such training should help employees understand the AI systems themselves, their rights in relation to said systems and their possibilities of redress, complaint etc. - AI should provide an opportunity for workers to apply their skills and competences to the fullest while at the same time remain owner of the production process. This includes the principle of transparency in the use of AI systems in HR, like the hiring of employees or the performance assessment of staff. It is important to safeguard the rights and freedoms of employees in line with non-discrimination rules as regards the processing of workers' data.- UNI Europa welcomes 5.1 – 5.4. We support that these examples raise real-life concerns of the adverse consequences of AI systems. - In 5.2. UNI Europa urges the group to expand on the issue of the human's right to know they are interacting with an AI identify. This	- In order to achieve "trustworthy AI", we need to establish public, independent and autonomous organisations that can control and audit (labour) algorithms (e.g. to identify underlying biases and the objectivity of data sets that train algorithms). Likewise, the implementation of the ethical guidelines on AI must be monitored. A European observatory focusing on the ethics in AI systems could play the role of an independent watchdog, including in business.- We would like the advice „to always keep record of the data that is fed to the AI systems" from the heading of data governance included under Accountability. For workers, it is paramount that the datasets used to evaluate performance, or in hiring or firing processes is transparent and can be accounted for. - The explanation of the principle of autonomy covers the question of AI at work only in a footnote, whereas this is an important issue that should be given a more prominent place. We would like to highlight the right of workers to individually and collectively opt out or withdraw from the use of AI systems (or a decision chosen by an AI system) if they undermine the workers' autonomy, decision making competence or disrespect fundamental rights and ethical principles. We recommend the inclusion of a special chapter that provides for ethical guidelines on AI in the work environment to address these issues more in detail.- UNI Europa ICTS welcomes that the HLEG on AI acknowledges the importance of social dialogue to realise trustworthy AI. We would like to add that the involvement of social partners, and in particular employee representatives, should not only take place regarding the general public debate on AI. Social partners should be involved in the establishment of codes of conducts, of standardisation schemes, development of training and in the proposed accountability governance. Employee participation and inclusion should take place early in the design, development and deployment of new technologies including AI and robotics. It is essential and important not only to inform and consult workers representatives in the work place or at branch level, but to enhance their co-determination rights and ensure their right to co-decide on the aims, reasons and	- We welcome that processes shall be examined in order "to allow a human control, if needed" (assessment list – governing AI autonomy). In so doing it should not be a question "to keep a human in the loop". We need clearly defined measures, that empower people to exert this control in all processes – regarding resources (technical equipment etc), organisational needs (time, liability, etc.) and qualification.- We suggest extending the list on the assessment of use cases (p.28) and add the question of processes, in order to use AI to ensure decent work (development and impact assessment).	- UNI Europa ICTS welcomes the call for Accountability Governance on page 21. The establishment of Data/AI Governance Councils in companies will indeed strengthen the accountability of AI systems and will address a weakness in the GDPR. The Council should consist of shop stewards and management and be responsible for holding management accountable and transparent to the use of AI and data. Whistleblowers should be able to address concerns to the Council and mandate the council to investigate on reported issues. - We welcome the process of developing guidelines for a trustworthy AI made in Europe, which encompasses corresponding "guidelines made in Europe", but would like to raise the question why non-European companies such as Google were granted full membership and full participatory right in the High- Level Expert Group. The status of associate expert would be more appropriate.- UNI Europa also supports the position of the ETUC regarding this consultation

					timeline and extent of changes in the labour market. The involvement of social partners is a must to find appropriate and future-proof solutions to concerns relating to employment, training, the nature of work, (in)equality or social systems and collective bargaining, especially at sectoral level.- As AI and automation have the potential to transform not only simple tasks but very complex processes, we need to have a large public discussion about the areas in which the use of AI is reasonable and beneficial for society. Part of the debate should be the question of how the profits generated by AI should be re-invested for the common good by creating employment in domains such as care, health services, education or mobility. Employees should participate in the distribution of profits, e.g. through wage increases or reduction of working time. Moreover, AI wins should be used to strengthen social security systems. This could be a measure to address the problem of future job losses and the precarisation of employment relations in a-typical work (e.g. platform work) due to AI and automation.- It is therefore important to integrate the aspect of the quality of jobs, decent work and social progress into the ethical approach in order to create a balance with the purely economic objectives underlying the creation and use of AI and robotics. - The Human-centric approach (HCD) not only presupposes information, transparency, participation and traceability, but also requires specific negotiation processes regarding decision-making in view of the aims and implementation of AI-systems at a very early stage for stakeholders such as employees and their codetermination bodies	could be done through a "labelling" system. For example, online bots should be labelled as such. Users should be made aware of the use of bots and AI in customer call-centre or help desks etc. - We would welcome that the employer-employee, employer-worker relation is explicitly mentioned in 5.3 as an example of power asymmetry. - Taking into account the power asymmetry in employer-worker relations, a separate point 3.6 on "workers' rights" should be added, which should contain the following points: "decent work by design", equal negotiation processes in the sense codetermination rights, informational self-determination of employees, non-discrimination principle and freedom of association including the right to strike. This is needed in order to secure worker's rights to co-decide on aims and application of AI systems, and create a legal framework.- Concerning the long-term risks and concerns we welcome that these should be considered. This could become an integral part of the accountability and audibility demands – i.e. that developers, users deployers etc need to reflect on the development/changing nature of the adopted AI as well as engage in predictions/forecasts of its future development scope and the consequences (positive and negative) hereof). - AI's influence does not only affect the world of work, but also democracy and society as a whole. We welcome that the draft refers to this point in Chapter I, paragraph 5.3., by stating that AI is not to be implemented in order to enable "citizen scoring" by a state/government. But this should also apply to private businesses. Neither states nor companies should be allowed nor have the possibility, to create	implementation of AI at the workplace. - Social partners at all levels should be involved in the implementation at company, industry, national or international level, including through collective agreements setting standards. In this context, it is required to describe the negotiation processes, e.g. central control structures for sector-specific solutions (cf. ‚AI Now' Report 2018): „Governments need to regulate AI by expanding the powers of sector-specific agencies to oversee, audit, and monitor these technologies by domain." (https://ainowinstitute.org/AI_Now_2018_Report.pdf)- Regarding the principles of accountability and transparency, we need to establish mechanisms for the protection of whistle-blowers who disclose the risks of AI systems or the non-respect of ethical principles – especially in the case of employees in companies that develop such systems. Internal reporting of risks and violations should be supported and rules in place to ensure follow up. - Organisations and companies should pay attention to potential biases encoded in the system development, training data and model performance – especially those that my affect the most vulnerable. They could also establish an internal ethical review process to democratise the decision-making process- Companies should not only increase transparency regarding the design and development of AI systems, but also in organisational chains of responsibility.
					human profiles such as "moral personality" or "ethical integrity". We reject the proposed opt-out-function and even possible "opt-in"-functions are not to be designed in a way that they conflict with fundamental human rights and possibly lead to the waiving of services that are useful for a person. AI-based services, that are important for work and life, must be designed in such a way that they do not require the collection of data which could be useable for human profiling.- Creating big data-bases always includes the risk of hackability as well as intentional and unintentional data-leaks. The guiding principle of "data-sovereignty" needs data-security in order to be viable. This implies explicitly not surveying data in areas that are of highly explosive nature for people in e.g. political, private or work-related areas. Fundamental rights as informational self-determination, the freedom of association and freedom of speech are not to be put at risk by creating such data-bases.		
Tom	DE SMEDT	Textgain			2.7. An interesting approach with regard to privacy is the use of stylometry, or AI that makes predictions based on writing style. Stylometry allows systems to predict a user's age, gender, region, state-of-mind, personality based on how they write, without even looking at content words (such as names of persons, places, etc.). For example, psychologist James Pennebaker found that men and women subtly use pronouns differently. Different age groups use punctuation marks and emoticons differently. Negative and positive personal opinions feature different adjectives, etc.	

					This also works on fairly small amounts of text, without the need to collect, store, and/or track users and their private information. Stylometry can be an interesting way to move forward in light of the GDPR. 2.10. A number of interesting new techniques are being discussed that are "model agnostic", i.e., they do not require the trained model to be cracked open. Rather, different variations of a user's input are tested to observe how these variations influence the outcome of the model. One interesting proposition in this regard are "counterfactuals" (see Sandra Wachter's work). The AI provides feedback about what needs to change in a user's input in order to reach a certain outcome (without actually revealing trade secrets of the AI etc.). For example, a user could ask: "Why did I not get a job interview?" If the system responds with "because you don't master the French language", then this is useful feedback for the user. If the system responds with "because you are not a man", then this is useful feedback for the developers and a warning sign that their training data is seriously biased.	
Max	Haarich	Embassy of the Republic of Užupis to Munich	The very welcome aim is formulated to not to provide "yet another list of core values and principlesfor AI". Later on, many of the usual supects for principles (e.g. accountability, transparency, safety, etc.) show up as "requirements" for trustworthy AI. The glossary should additionally discriminate rights, values, principles and requirements. These terms are the necessary layers to translate an often implicit, abstract ethical mindset via multiple layers into precise actions. Yet, these and many other "layer-terms" are used differently and contradictorily in the public debate. If there was one key player, like the HLEG, to define these terms in an instrumental way and disseminate these definitions broadly, ethical frameworks would become more connective and the whole European and global debate would benefit.	It is a great idea to root the considerations in already existing human rights to have one focal point of orientation.	One of the most efficient methods to secure ethically aligned ouputs is not mentioned here: including an artistic perspective into the technological development process. The EU funds such projects within its program "S+T+ARTS" with the confidence that arts can make AI-applications more inspired and more human-centric, i. e. more ethical. Arts and culture offer centuries of ethical discussion about artificial kinds of intelligence dating back at least to the Golem figure in the Jewish Talmud and culminating into thousands of sci-fi-movies since Capek's R.U.R from the 1920s. If ethics shall guide the decision on what to do, arts, culture and especially sci-fi already offer and discuss many alternatives. Additionally, a manyfold of artistic and cultureal projects dealing with ethics in AI already exists and they are rooted in the very heart of civil society. They can help translate the ongoing developments and create broader understanding of the technology leading to more effective adoption of AI on societal level.	The assessment list is a great idea and gives deeper understanding of the single requirements.	We appreciate the design of ethics for AI as an open and iterative process. In hardly any other AI strategy this can be found this succintly formulated. Yet, if you want to motivate precise and constructive feedback, you could offer some guiding questions on the most debatable points from your own perspective.
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential

Anonymous	Anonymous	Anonymous	If trustworthy AI is our northern star, we should provide a clear definition of trust here. A good starting point may be that of Lee and See (John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. Human Factors, vol. 46, pages 50–80, 2004.): "Trust can be defined as the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability." Muir and Moray provide a definition of trust tailored to automation that might well be adapted to AI (Bonnie M. Muir and Neville Moray. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics, vol. 39, no. 3, pages 429–460, 1996.): "[Trust] in automation is a composite expectation of (1) the operator’s general expectation of the persistence of the natural physical order, the natural biological order and the moral social order, (2) a specific expectation of the technical competence of the automation and (3) a specific expectation of the fiduciary responsibility of the automation." Furthermore, the introduction should be a little bit more electrifying: At the moment there is a lot about ensuring that basic principles of society and democracy remain intact. While this is absolutely necessary, something else needs to be emphasized too: What do we actually want to ACHIEVE? I think if we employ AI in the right way, we can help master some of humankind's biggest challenges: hunger, diseases, climate change, you name it. And this is what the European AI community should strive to do. This is what human-centric AI essentially means. We should communicate this more.	4. Ethical Principles in the Context of AI and Correlating Values I would add here the Principle of Acceptability as discussed in this article: https://www-docs.b-tu.de/fg-technikwissenschaft/public/BTU_News_09_2018_Die_Ethik_der_k%C3%BCnstlichen_Intelligenz.pdf (sorry, only in German). It essentially says that humans should at all times be able to position themselves in favor or against an AI system. Three conditions must be met to guarantee this: (1) It must be transparent to a human whether they are interacting with an AI or not. (2) The processes in an AI system must be made comprehensible to the human on request. (3) A broad public discussion must be fostered in order to decide which systems and developments are considered useful and desirable and which are not. 5.4 Lethal Autonomous Weapon Systems (LAWS) If we take our approach "human-centric AI" seriously, it should be crystal clear to us that AI must never be in a position to take control over human lives. Hence, the EU should ban all LAWS. 5.5 Potential longer-term concerns The possible consequences of the successful development of an AGI (however improbable this event might seem) might be disastrous to humankind. The argument that we only need to make it safe is not valid: An AGI will most certainly be able to change its utility function; otherwise it would not be an AGI. This means that whatever we incorporate into its utility function (e.g., the do no harm principle, pg. 9) can be overwritten. And even if we somehow manage to construct an AGI that cannot change its utility function and is benevolent towards humankind it may interpret this in a different way than we would (see, e.g., Thomas Metzinger's BAAN scenario for an example: https://www.edge.org/conversation/thomas_metzinger-benevolent-artificial-anti-natalism-baan). The emergence of AGI might have the effect of an atomic first strike against humankind (and if it does, it will not matter whether European, Asian, American, ... researchers have created this AGI). Therefore, the EU should not fund research on AGI or even research that accepts the risk of creating AGI as a side-effect. On the contrary, our goal should be to ban this kind of research on an international level.	2. Technical and Non-Technical Methods to achieve Trustworthy AI Traceability & Auditability In my opinion, this section is too vague. We should strive to develop safety regulations on different levels that AI systems must conform to. As an example safety regulations from the automotive/avionics industries can be considered, e.g. safety integrity levels (SIL) or ISO 26262.	I would be glad to contribute further. Do not hesitate to get in touch.

Andreea	GULACSI-GOLOGAN	CEN and CENELEC	This document aims at providing a response of the European standardization organisations CEN (www.cen.eu) and CENELEC (www.cenelec.eu) to the first draft of the "AI Ethics Guideline", produced by the High-Level Expert Group on Artificial Intelligence (AI HLEG).Artificial Intelligence constitutes indeed one of the most important issues for the future of the industry, as AI applies to a variety of sectors where standardization is of high relevance in and is set to have a powerful impact on how businesses operate as well as on the way the job market is currently structured. In the context of AI-enabled products, the European Commission has already foreseen impacts in several sectors with changes or clearer guidance brought to a series of EU Directives (such as RED, Machinery, Product Liability).Recognizing the increased importance of AI to a variety of products, technologist and services , CEN and CENELEC have decided to establish a Focus Group on AI. The group will be in charge of addressing the challenges identified in the EC Communication COM (2018) 237 referring to the deployment, interoperability, scalability, societal acceptability/concerns, safety and liability of AI as well as identifying special European needs, which are not taken care of in the framework of the international standardization organisation ISO/IEC JTC 1 SC 42 (on Artificial intelligence). One of the main objectives of the group will also be to analyze the ethical aspects that need to be tackled in relation to AI and to investigate the possibility and needs for establishing CEN-CENELEC's own set of "AI ethical guidelines", as suggested within the Stakeholders' Workshop on "Trustworthy AI – building a framework with standardization", organized last September 2018. Other current relevant and complimentary standardization initiatives addressing AI at European and international level, which are worth mentioning, are:- International standards 1) within IEC with the group SEG 10 focusing on Ethics in Autonomous and Artificial Intelligence Applications (IEC - SEG 10: Ethics in Autonomous and Artificial Intelligence Applications > Scope); 2), more broadly within ISO/IEC JTC 1/SC 42 Artificial intelligence (https://www.iso.org/committee/6794475.html).- COPOLCO - ISO's Committee on Consumer Policy (https://www.iso.org/copolco.html) is undertaking work to articulate the ethical approach from a standards and consumer perspective. An initial paper has already received positive feedback and will be further developed in the coming months.OCEANIS (Open Community for Ethics in Autonomous and Intelligent Systems- https://ethicsstandards.org/), which includes several CEN and CENELEC members (BSI UK, DKE Germany, ASI Austria, NSAI Ireland). OCEANIS is a global forum for discussion, debate and collaboration for organizations interested in the development and use of standards to further the development of autonomous and intelligent systems.I. Rationale and foresight of the guidelines As stated within the document, the guidelines: "should be seen as a living document that needs to be regularly updated over time to ensure continuous relevance as the technology and our knowledge thereof evolves". The	In this chapter experts address the issue of fundamental rights, developing a set of principles and values that underpin ethical purposes for AI. A special focus is dedicated to transparency within the AI decision-making process, demanding AI systems to be auditable and explainable. CEN-CENELEC strongly endorse these points as they are indeed amongst their constituent principles as key factors for a functioning and trustworthy market economy.	In this second chapter experts propose both technical and non-technical methods that can serve to help realising and implementing Trustworthy AI. Within the "Non-Technical Methods", standardization is among the suggested approaches. CEN-CENELEC recommend this, as standards are designed to deliver clear and unambiguous provisions and objectives. Standards are voluntary and separate from legal and regulatory systems, however they can successfully be used to support or complement legislation. Standards are developed when there is a defined market need, through consultation with stakeholders and a rigorous development process. When the European Commission released its strategy on AI through COM(2018) 237 on 'Artificial Intelligence for Europe' on 25 April 2018, it underlined how a "solid basis of standards" already existed in the area of AI-enabled devices. The Communication recognised as well that "further development and promotion of such safety standards and support in EU and international standardisation organisations will help enable European businesses to benefit from a competitive advantage, and increase consumer trust".Concrete proposed text to add in the Draft Ethics Guidelines on AI p. 21."International standards from either ISO or European standards from CEN allow for a transparent set of requirements or guidelines (as is the case with ISO 27000 on it-security) that can help consumers, producers, legislators etc. specify requirements for the AI products they want to buy, use or sell. The use of standards can contribute to establishing minimum criteria and thus trustworthiness and trust around the products or systems. Without established standards it is difficult to compare and evaluate the products and systems for non-experts. Standards can be used for self declaration or for certification, like the European CE mark, depending the market need.Private labels like Fairtrade can be difficult to evaluate against each other and can lead to fragmenting the market."	Within the last chapter, the document provides a list of assessments, with the aim of helping operationalise the achievement of a Trustworthy AI. The list is not proposed as conclusive and CEN-CENELEC will certainly take on themselves to potentially review or further improve it in the future, with the contribution of their members and other relevant stakeholders.	CEN and CENELEC are aware of the fact that a shared definition of ethical standards will be instrumental to successfully overcome the challenges raised by the use of gradually more powerful AI tools within European as well as within international business operations. Hence, CEN and CENELEC positively receive this first draft of ethical guidelines and appreciate the transparent use of consultation with all interested stakeholders. CEN and CENELEC offer their support by putting forward the role European and international standards can play for the future market uptake and deployment of AI. Moreover, CEN and CENELEC receive favourably and support the decision of the High-Level Expert Group to include standards within the "Non-Technical Methods" defined in the second chapter of the document and hope to see it laid out in the final version with the same consistency.The draft document also foresees that "mechanism will be put in place that enables all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis". CEN and CENELEC believe that the certification of those mechanisms should be based on relevant and reliable standards.As a final point, further focus should be dedicated to the potential development of voluntary standards, which have demonstrated to offer a flexible, adaptive and collaborative alternative to regulation by providing common languages, terminologies, guidelines and good practice developed by and for stakeholders. CEN and CENELEC's robust standards development process requires open and full consultation with stakeholders to build consensus-based outcomes. This gives standards the legitimacy and degree of market acceptance useful for public policy purposes.

document contains a glossary, which is in line with CEN-CENELEC approach as the importance of terminology is recognized in standardization, both as part of general as well as of specific standards, with specific terms and definitions. A common understanding of terms is needed in order to clarify those aspects currently under analysis and later “build out” on other, new areas. CEN and CENELEC recommend that the European High-Level Expert Group on AI engages directly with the international standard works under development, covering this specific topic (ISO/IEC WD 22989 Artificial intelligence - Concepts and terminology - https://www.iso.org/standard/74296.html?browse=tc) to ensure there is a universal and consistent definition of key terms. CEN-CENELEC also suggest the European Commission includes within the final document the results of the ongoing work in the ISO/IEC JTC 1, SC 42 on AI, which is currently laying down the foundation for future AI standardizations, starting with the development of a standardized common terminology as well as technical reports on AI use cases and characteristics of artificial intelligence systems. Further standardization work to be taken into account are the findings of ISO/IEC JTC 1 SC 42, which also includes a Working Group on Big Data.Within the document, the need to set up a “mechanism… that enables all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis” is also recognized: this embraces CEN-CENELEC defined ambition to have transparent and inclusive access in the development of standards and regulations. Moreover, CEN-CENELEC propose the use of European standards, connected wherever possible, through long-standing robust mechanism, to established international standards. Via the ordinary CEN-CENELEC processes, the guidelines have the potential to be regularly updated, taking into account the views of all relevant stakeholders (consumers, SMEs, governments, etc.) from across Europe. Certification and accreditation are not within the competences of CEN-CENELEC, however there are established mechanisms in place to ensure consistency and quality of certification and accreditation. CEN-CENELEC recommend the European High-Level Expert Group on AI engages with European Accreditation (https://european-accreditation.org/) and IIOC (Independent International Organisation for Certification - http://www.iioc.org/) to establish a reliable certification framework. It would be useful to make a clear distinction throughout the document between principles, governance and management. Governance relates to how a company should address AI related questions and challenges, opposing to how to answer them. Management are the engineering techniques used to implement AI systems. By using such distinctions and taking care of not exercising judgement, a way is given to defuse fear and increase the societal acceptability of AI.

Clemens	Otte	BDI - Federation of German Industries	Rationale: The BDI welcomes the approach of the High-Level Expert Group (HLEG) to define ethic guidelines for trustworthy AI. To foster support for the guidelines and facilitate their impact in common practice of AI development, deployment and use, the need/reasons for AI guidelines should be better motivated. The HLEG should carve out that the development of AI based on European ethical and societal values can build trust in artificial intelligence, facilitate a broader uptake of AI and can serve as a unique selling proposition for AI "made in Europe".Aim/Scope: It is positive that the guidelines aim to support companies in implementing ethical principles and values in the development and application of AI. Although further adjustment is needed. The guidelines do not yet take sufficient account of the fact that the ethical boundary conditions of AI systems differ considerably depending on the field of application. Particularly for industrial applications ethical questions often play only a minor or very context-specific role. An insufficient differentiation regarding the criticality of AI applications may lead to undifferentiated red lines, which unnecessarily restricts Europe's competitiveness.Endorsement mechanism: BDI appreciates the target of putting a method in place to enable all stakeholders to formally endorse and sign up to the guidelines on a voluntary basis. This can support transparency for users and foster trust. Due to the diversity of AI applications, however, a "one size fits all"-solution seems not to be feasible. The suggested technical (and non-technical) methods (as mentioned in Chapter 2) and detailed checklists (as described in Chapter 3) are too specific and not applicable for all use cases. Thus, a "holistic" framework including high-level principles, seem more appropriate for the commitment/endorsement proposal of the HLEG. The methods and checklists should not be included into the formal endorsement but added as best-practice examples instead. Furthermore, the endorsement process raises questions regarding its practicality. The guidelines do not make clear what consequences an endorsement has on the signatories, e.g. if signatories thereby fall under specific external governance or auditing.Glossary: The definition of AI seems not to be appropriate for a political debate, since it does not differ between "weak" and "strong" AI. The choice of words such as "perceiving" or "reasoning" falsely suggests that AI systems are human-like, fully autonomous, thus potentially prejudiced, acting systems. Furthermore, the formulation "… and deciding the best action(s) to take to achieve the goal" should be replaced by "… and providing predictions or results, which might be implemented automatically where appropriate and traceable."Moreover, a definition of ethics should be included in the glossary, since the perception of ethics differs in different cultures.	Definition of Ethical Purpose: BDI values the approach to derive responsible/trustworthy AI development, deployment and use of AI based on fundamental rights, ethical principles and values. But the guidelines should carve out more precisely the EU understanding of terms like "ethical purpose" and "wellbeing and the common good" (p. 5). The wording may indicate, that any AI should serve a higher ethical purpose or even the sole purpose of AI should be ethics.Equality, non-discrimination and solidarity including the rights of persons belonging to minorities: AI development, deployment and use should adhere to the fundamental right of equal treatment, as set out e.g. in Chapter III of the EU Charter of Fundamental Rights. In this context, the document states that "equality of human beings goes beyond non-discrimination" (p. 7). To avoid misconceptions, conflicts with the fundamental right of individual freedom, or the notion of a 'levelling down', this statement should be revised to capture more precisely what it is supposed to mean in the context of AI.Informed consent: The concept of an „informed consent" needs clarification. It remains unclear whether an "informed consent" is identical with the consent under the GDPR. It should be made clear that only the term used in the GDPR applies. Moreover, the GDPR offers further legal basis for data processing, such as processing necessary for the performance of a contract or for legitimate interest. Therefore, the notion of informed consent is given too much prominence and misleads it to be the only and best requirement to preserve autonomy.Comments regarding the Ethical Principles and Values: The BDI agrees with the five principles and correlated values in general. However, some adjustments are necessary:The Principle of Beneficence "Do Good"Taking into account that AI should create added value to different stakeholders, economic interests have to be considered as legitimate interests of a company in order to promote economic growth. Adhering to principles such as traceability, transparency and self-determination might come to an economic cost. Therefore, the principle of beneficence should be complemented by a notion of proportionality.The Principle of Non-maleficence: "Do no Harm" BDI agrees, that AI systems should protect the dignity, integrity, liberty, privacy, safety, and security of human beings. AI applications are being developed by humans, and while it is understandable and correct, that the expectations are higher than towards humans, it must be understood, that high efforts and continuing improvements are necessary to reduce potential risks. Use case specific, the necessary quality level and fault tolerance, fall back solutions in case of error, necessary effort for testing, monitoring/controlling have to be defined, always under consideration of the field of application, how autonomous the AI may act, the opportunities and possible risks, and which machine learning method is used. Moreover, BDI supports the aim to avoid discrimination, manipulation or negative profiling in general. However, "negative profiling" is necessary in a certain context. Businesses must be able to segment customers and business partners based on certain predefined criteria to assess the creditworthiness of a person by using AI. This is also essential for insurances, financial	Accountability: The HLEG rightly points out that the choice of accountability mechanisms is highly dependable on the use case, the field of application, the autonomy and many more factors. Regarding accountability, a highly differentiated approach should be targeted. Data Governance: The statement that "the datasets gathered inevitably contain biases, and one has to be able to prune these away before engaging in training" depends on the aim of a given policy or algorithmic model. Pruning a model to make it fairer for one group may inevitably create biases and unfairness for another group, in particular if different groups have different descriptive distributions and base rates. It thus makes more sense to identify bias / unfairness with data that reflects the real, imperfect world and then correct post-processing for bias and unfairness (which often would be relevant for minority groups in a machine learning setting).Design for all: The requirements for „design for all" are far too general and not applicable to all AI solutions, especially with industrial applications. For example, AI included into automated cars (e.g. level 3) could still not be used by people without driver license, thus a use of the service/product will not be available for all ages. Furthermore, it is highly likely that there will be AI-based products and services that appeal to particular groups rather than universally to all humans, e.g., gender specific apps, age specific apps (and combinations thereof). Governance of AI Autonomy: As mentioned by the AI HLEG, the use cases and the fields of application differ, thus the impact (benefit and risk) also differs immensely. Due to this fact, BDI proposes to always review what level of autonomy in decision should be applied (AI only as a source of information, AI as an assistant with final decision by user or AI acts fully automated without human involvement). Furthermore, it is essential to also review the level of autonomy in learning (may the AI learn on the market (retraining possible), with limited parameters (no safety relevant parameters) or no learning/evolvement on the market possible). And as a third dimension, the level of risk should be considered (e.g. which persons or laws could be harmed and how). Such a structure could be very helpful to take a more differentiated view of ethical issues.Transparency: Different Levels of Transparency will be necessary for different use cases and groups. The right level of transparency or explainability is important to strengthen the user's trust in AI applications. However, higher transparency will be necessary for developers and operators to ensure quality monitoring and continuous improvement. Furthermore, for use cases with potentially higher risks, higher levels of transparency are necessary. (Non-)technical methods: As mentioned above, the proposed methods are not applicable for all use cases. Thus, they should only be considered as best practice and should not be included into the formal endorsement. Additionally, as a non-technical method, all AI systems should come with a clear description of their limits, including the areas they are intended for and those, they are not intended for, as well as description of input data that the system cannot properly cope with (e.g. a system tailored to autonomous car control might not properly cope with autonomous truck control	BDI welcomes the efforts of the HLEG to offer guidance to steer developers, deployers and other innovators toward ethical purpose and technical robustness. However, the list of the HLEG is very inconsistent. The questions vary in their granularity and do not differentiate between the AI methods being used. The questions are not suitable as practical assistance yet, since they lack technical details and specification (which is crucial for real guidance).	Costs of implementing trustworthy AI: To implement AI successfully in Europe it must safeguarded that additional costs and bureaucracy are minimized, e.g. the additional auditing services and storage of logs would demand additional development effort, operational cost for storage, processing power, licenses, as well as man power to monitor and maintain the auditing the system. The costs can be a high burden, especially for small and medium-sized enterprises.Alignment with existing processes: The specific recommendations and guidelines should be more aligned to existing processes for data protection, product safety and security. Aligning it would mean to make a fit/gap analysis with existing procedures and integrate it into them in a lean manner. Within this context, the guidelines should point out that AI is not a completely new technology. Many industry companies have long term experiences with AI and already established well-functioning processes to minimise the risks. This might also help to calibrate societal concerns, presenting AI not only as disruptive and transformational but also as an incrementally developing technology.Fostering R&I: Fostering R&I on achieving Trustworthy AI in EU should get a prominent position in the document, at the forefront of activities, including practical test cases (e.g. sandboxes) for various verticals.Support companies to operationalise the guidelines: Companies will be responsible for operationalising these guidelines. A path to establishing measures for these companies should be described in detail.

institutions and e-commerce businesses.The Principle of Autonomy: "Preserve Human Agency"A general "right to decide to be subject to direct or indirect AI decision making" is impractical. A differentiation with regard to criticality and context is urgently necessary. According to the wording, every citizen could have the right to object to an uncritical use of AI in a longer process chain by which he or she is indirectly impacted (e.g. if an AI is used to calculate the time of garbage collection from a household). A general, non-context-specific right to opt-out would be highly impractical and would hinder the uptake of AI in administration and business processes. It should also be considered that there are limits to the right to be subject to direct or indirect AI decision making. Suitable alternatives are not available in all cases. This applies in particular to the working environment mentioned in footnote 13. The formulation "…anyone using AI as part of his/her employment enjoys protection for maintaining their own decision making capabilities and is not constrained by the use of an AI system" should not be interpreted as an individual right to object to any AI implementation in the working environment. Today, AI is already an inherent part of the working environment for many professions (e.g. pilots are supported by AI in aircrafts). Employees should participate collectively in decisions around the implementation of AI systems in working environments through established bodies of representation.The Principle of Justice: "Be Fair" Instead of stressing "that AI systems must provide users with effective redress if harm occurs", the guidelines should emphasize that ultimately humans are responsible. Operators of AI should know and make clear who is responsible for which AI system or feature. As with other technologies and products, the people who design and deploy AI systems must be accountable for how their systems operate.The Principle of Explicability: "Operate transparently" BDI fully agrees with the AI HLEG, that explicability is a key success factor to increase the acceptance and trust in AI systems. For this, it is important to explain the function of AI in an understandable manner. Focusing on explaining the result, the base for decision making and the benefit of the system seems to be the key. In order to achieve a high explicability, non-AI specialists could be involved in the design process.However, transparency, especially when using deep learning, still has its limits. But the limitations (e.g. accuracy, safety), the decision process, the algorithm and the defined quality criteria should be made transparent and documented within companies. Additionally, it needs to be clarified how potential neutral audits in critical contexts could be ensured, especially considering the limited pool of AI specialists.Potential longer-term concerns: The probability of potential occurrences as mentioned by the HLEG are currently very low and well into the future. Therefore, we suggest focusing on realistic and existing challenges but remain attentive to future development of critical topics. However, Artificial Moral Agents (AMAs) should not per se pose a threat as long as these have been trained within a given and acceptable ethical framework. It is highly likely that AMAs being trained by re-enforcement principles due to different dimensions and requirements). Moreover, another method/paragraph should be added on school education, vocational training, required curricula & skills for the new and working generations.

(where the reward is adherence to the ethical principles) are near-future feasible (i.e., white swans). This is decidedly not a negative development and might be one of the few technology principles existing today that might actually work in terms of developing practical ethical AI.

Magdalena	Piech	European Tech Alliance			AI, often called machine-to-machine learning, is in essence a discipline that uses digital technologies to generate systems able to autonomously reproduce human cognitive functions, especially the apprehension of data. In other words, AI is a discipline that improves human capability of learning from very large data sets and often generates tools that replace the human being when making decisions based on a vast amount of data. As any discipline, the potentialities of AI can be infinite, which is why any AI tool must carry society's common values or, as the AI HLEG recommends, it must be human-centric. The AI discipline cannot ignore that societies organise themselves through a set of institutions, operating according to the rules that reflect the common values of the community of citizens which has decided upon them.Consequently, the "Principle of Beneficence: Do Good" and the "Principle of Non- maleficence: Do no Harm" should be inherent to AI design. We believe that the pre-conditions to implementing these two principles are: - to raise awareness among researchers, developers, and decision makers that they are bound by the same set of values and rules as the community of citizens, therefore having to fight against the same risks of replicating biases through AI and be cautious that their research results are not being misused.- to help identify a methodology to eliminate those risks and enable the AI research community to assess AI innovations from all possible angles (e.g. legal boundaries, economic value, social and societal implications) As AI requires large amounts of data for learning purposes, the data should be carefully selected and be relevant to the AI's objective. This, in our view, defines the "Principle of Justice: Be fair".The other facet of an ethical AI development should be the "Principle of Explicability: Operate Transparently", meaning that the purposes of the uses of AI tools must be clearly demonstrated, by using non-technical language.			The European Tech Alliance (EUTA) represents leading European tech scale-ups that were successfully built in Europe. We are the sole organisation that represents home-grown European tech companies from across the whole industry. Therefore, we demonstrate the variety of European tech voices and business models, ranging from digital music services to big data search, e-commerce platforms, mobile games, carpooling, and file sharing.The EUTA would like to commend the initial work of the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG). The draft guidelines showcase a human-centric focus from the group's experts, as well as a welcome practical approach to move this debate forward. The EUTA supports the EU's ambition to be deliberate in creating a unique approach to AI based on EU norms and values. Such an approach is a necessary stepping stone as the AI discipline is progressively building value in all sectors of human activity by using technologies that can replace human cognition to some extent. Concerns regarding how one can trust AI are natural, and need be broken down by imagining and implementing a principle-based AI governance model. In the ever-pressing global AI race, various regions are investing to be the 'first' to reach the finish line. However, we do not believe that being 'first' necessarily means being a leader, and we recognise that a European approach that prioritizes responsible and trustworthy AI and which responds to citizens' needs gives us a greater chance of long term success. While the AI HLEG draft has been specifically tasked with the mission to identify a set of ethics guidelines, we must not forget the context in which we look to progress. On a macro level, EU investment in AI research and development severely lags behind the US and China. Moving forward, we need to ensure that we maintain our future framework of values whilst not risking the opportunity to become globally competitive, using the ability to innovate and experiment with large datasets. We appreciate the risk analysis that the authors of the draft report have recognised and touched upon in their work. The field of AI is developing at a rapid pace and over-regulation will chill its development. Future policies in the field should take account of the intricacies of each industrial sector and the likelihood of the adverse consequences of over-prescriptive expectations. For policies to promote AI that is truly European, they must respond to people's needs. European tech firms are amongst those who have, and will continue to lead in investing in its development, creating jobs and investing in skills.For example, Criteo's AI Lab in Paris is pioneering computational advertising based on transparency and user control. By promoting open research methods, the Criteo AI Lab's experts will enable the AI research community to power new AI applications.Clear guidelines that promote innovation and entrepreneurship are needed to help us all navigate the technological developments possible through AI. The precursor to our industries' success will be a positive ecosystem that focuses on early stage development across a range of sectors, rather than the lone establishment of overly-restrictive barriers.We look forward to the AI HLEG's final publication and are ready to contribute to the ongoing discourse

							surrounding AI. The EUTA's members are uniquely placed to advise and showcase the real world needs and capabilities of AI in the EU tech ecosystem.FURTHER COMMENTSThe vast scale of AI investment occurring outside the EU should be a wake-up call for policy makers across the European Union. We must ensure that the EU becomes a leading player in guiding EU principles based on European values, meaning:• Creating a level playing field between European and Non-EU AI-driven companiesThe ability to innovate and experiment with large data sets is key to providing state-of-the art AI technologies that help solve society's most pressing problems and meet consumers' needs: - For authorities to identify early signs of natural changes, an AI system collects vast amounts of data on weather conditions, agricultural yields, historical land data, social media information and so on;- For retailers to tailor an offer of products and services convenient to new markets, an AI solution relies on multiannual shopping trends data, consumers' recommendations, sellers' inputs and so on.Today, a small number of dominant non-European players have the capacity to continuously improve their forward-looking AI-based services based on the collection of such data. The EU needs to take a holistic approach to ensure Europe's AI leadership is given the platform to compete on equal terms against these non-EU players.• Curbing EU's 'brain drain' in AIToday, there are too few examples of young researchers demonstrating skills in both informatics and mathematics and emerging to meet the growing number of opportunities in the AI sector. In the talent race, the EU prepares the leading expertise to shape the AI ecosystem. However, in the end, non-EU players have the necessary resources to attract the most skilled individuals. We must ensure that more is done to identify solutions to make sure that our talents add value to the EU economy. We therefore present questions to open the discussion on curbing the EU's brain drain in AI: - How can we accelerate the emergence of AI training modules across the EU?- How do we encourage talents across the world to train and work in the EU?- Should the EU 'put a price' on the AI talent competition that would dissuade dominant players from absorbing the talent trained in EU schools (e.g. financing studies, establishing 'transfer fees' when hiring from the competition)?
Maria	Reiffenstein	Federal Ministry of Labour, Social Affairs, Health and Consumer Protection	Künstliche Intelligenz ist in ihren komplexen technologischen Möglichkeiten und vielschichtigen Auswirkungen von besonderer Bedeutung und reicht in alle Teile der Gesellschaft hinein. Sie bedarf daher gründlicher Diskussion auf allen Ebenen und berührt auch die essentielle Frage des Menschenbildes bis hin zum Stellenwert des Bewusstseins. Es wird daher sehr begrüßt, dass die EK eine HL-Expert Group beauftragt hat, ethische Guidelines und Empfehlungen für die Politik zu erstellen. Der normative Charakter der Guidelines ist allerdings nicht ganz klar. Die ethischen Guidelines sollen – so der Entwurf – von allen Stakeholdern auf freiwilliger Basis unterzeichnet und damit „bestätigt" werden. Laut Guidelines werden auch öffentliche Institutionen als Stakeholder bezeichnet,	In I.2 wird der Grundsatz der Autonomie als Wert beschrieben, der unmittelbar in den Menschenrechten zum Ausdruck kommt. Im Anschluss wird skizziert, dass das Individuum frei und vernünftig entscheiden wird, wenn es nur ausreichend Informationen hat. Aus der Behavioural Economics Theorie wissen wir, dass diese Sichtweise sehr idealistisch und nur teilweise richtig ist. Die in der Folge genannten Grundprinzipien (1.3) werden als essentiell unterstützt. Die angeführten konkreten Beispiele erwähnen allerdings Konsumenten, die wesentlich Adressaten von KI sind, nicht adäquat. So wird bei der Freiheit des Individuums zwar der Aspekt der Befähigung zur Kontrolle über das eigene Leben genannt. Dies solle – so die Guidelines - auch die Freiheit zum Unternehmertum, zur Wissenschaft oder zur	Die angeführten Erfordernisse wie auch die Methoden der Umsetzung werden grundsätzlich unterstützt. Die Ausführungen sind allerdings sehr verkürzt. So konzentrieren sich die Aussagen zu notwendiger Regulierung auf Sicherheit, Haftung und Abhilfemechanismen. Für die Sicherstellung der Entscheidungssouveränität und der dafür notwendigen Transparenz wird offenbar keine rechtliche Regulierung für notwendig erachtet. Auch die Konsumentenbildung wird als Domäne von Ethikern ohne rechtliche Dimension beschrieben. Die Asymmetrie zwischen Herstellern/Anbietern und Konsumenten im Bereich KI ist aber wesentlich größer und kann viel schwerer durchschaut werden als im klassischen Geschäftsleben. Wenn man die im ersten Kapitel skizzierten Grundrechte	Die gestellten Fragen sind sicherlich nützlich, aber natürlich bei weitem nicht vollständig. So stellt sich in III.4 bzw I.7 die entscheidende Frage, ob die Systeme in geeigneter Art und Weise ermöglichen, dass Konsumenten – abgestuft nach verschiedenen Funktionen – differenziert (und nicht nur pauschal) zustimmen oder ablehnen können (privacy by design). In I.7 wird die Möglichkeit des Users, Algorithmen zu erfragen, als notwendig angesehen. Dass die Informationen über das System des Algorithmus' einschließlich seiner Auswirkungen und wichtigsten Parameter eine Bringschuld des Anbieters ist, wird offenbar nicht als wichtig angesehen. Diese Informationen sollten aber nicht nur auf Nachfrage gegeben werden. Unklar bleibt, wer jeweils Adressat dieser	

sodass sich auch die Frage stellt, ob intendiert ist, dass auch Behörden oder (wohl eher nicht) auch die EU Organe die Guidelines unterzeichnen. Weiters stellt sich die Frage, welchen Stellenwert bzw welche Konsequenzen die Unterzeichnung hat. Begrüßt wird, dass die Guidelines als „Starting point" und lebendes Dokument bezeichnet werden.
Das Verhältnis von Ethik und Recht wird in den Guidelines unterschiedlich beschrieben. Einerseits werden die Grundrechte als die Basis der ethischen Prinzipien bezeichnet. Andererseits sollen Grundrechte, ethische Prinzipien und Werte zusammen den „ethischen purpose" sicherstellen. In welchem Verhältnis Ethik und rechtliche Regulierung zueinander gesehen werden bzw in welchem Ausmaß ein Bedarf an (über Grundrechte hinausgehender) rechtlicher Regulierung gesehen wird und wer diesen bestimmen soll, bleibt offen. Zu Beginn heißt es nur, dass die Guidelines rechtliche Regulierung nicht ersetzen sollen. Im 2. Kapitel, das sich mit der Implementierung der Guidelines beschäftigt, ist dies nur sehr allgemein angesprochen. Der koordinierte Plan für künstliche Intelligenz (COM(2018)795 von 7.12.2018) nennt unter den notwendigen Maßnahmen auch die Überprüfung des rechtlichen Rahmens. Wichtig wäre es, den rechtlichen Rahmen in Abstimmung mit den ethischen Guidelines zu prüfen. Während der Koordinierte Plan auf S.5 das Mandat der HL-Expert Group so beschreibt, dass neben ethischen Leitlinien auch Empfehlungen an die Politik für ua. den notwendigen Rechtsrahmen erstattet werden sollen, lautet das Mandet der HL-Expert Group in den Ethics Guidelines lediglich die Guidelines und „policy and investment recommendations". Unklar ist somit, ob die HL-Expert Group auch den geeigneten/notwendigen Rechtsrahmen vorschlagen soll.
Auch die Entschließung des Parlaments von Februar 2017 (P8_TA(2017/0051) ist zu erwähnen, die wichtige Empfehlungen an die EK betreffend zivilrechtliche Regelungen im Bereich Robotik formuliert. Darin wird beispielsweise auch eine Registrierung intelligenter Roboter thematisiert, um gezielt die sensiblen Anwendungen im Auge zu behalten und bei unerwünschten Entwicklungen notfalls gegensteuern zu können.

Kunst einschließen. Die Konsumentenrolle wird mit keinem Wort erwähnt.
Konsumenten werden lediglich in I.3.4 (Gleichheit, Nichtdiskriminierung und Solidarität einschl. Minderheiten) im letzten Satz, aber ohne sinnvollen Zusammenhang angeführt (Equality also requires adequate respect of inclusion of minorities, traditionally excluded, especially workers and consumers). Was hier mit "especially" gemeint ist, ist nicht verständlich. I.3.5 weist auf die Bedeutung der Bürgerrechte hin. So müssen Bürger informiert sein, wenn ihre Daten automatisiert verarbeitet werden und auch ein Recht haben, dies zu verneinen (opt-out). Auch im Zusammenhang mit dem Prinzip der Autonomie (I.4) wird das Recht auf opt-out betont, obwohl gleichzeitig dem Konsumenten das Recht zugestanden wird, dass er selbst entscheiden können muss, ob er von KI-Entscheidungen betroffen ist. Dazu würde aber eine Vorwegentscheidung (opt-in) notwendig sein. Die weiteren in I.4 formulierten Prinzipien werden begrüßt. Insbesondere das Prinzip der „Explicability" ist aus Konsumentensicht entscheidend und impliziert jene Transparenz, die es jedem ermöglichen muss, das Funktionieren von KI zu verstehen (s.S.10).
Zu I.5.3 Normative & Mass Citizen Scoring: Scoring wird nur im Zusammenhang mit Bürgerrechten als kritisch angesehen. Entsprechend wird die Schaffung von opt-out-Rechten gefordert. KI-Algorithmen, die Konsumenten in verschiedensten Bereichen des Geschäftslebens bewerten und damit auch ihre wirtschaftlichen Möglichkeiten unter Umständen massiv einschränken, werden bedauerlicherweise nicht erwähnt.

und Prinzipien ernst nimmt, müssen diese auch auf allen Ebenen umgesetzt werden. Die Frage des Rechtsrahmens wird dabei eine entscheidende Rolle spielen.

Fragen ist.

Franca	Salis Madinier	CFDT Cadres	- CFDT cadres welcomes the possibility to contribute to the stakeholders' consultation and underlines the importance of a broad public debate and information on AI. This debate must result in clear ethical and social guidelines and standards with the aim of improving the living and working conditions of European citizens.- We acknowledge the innovative potential of AI and new technologies that can be beneficial for our society. However, these new technologies also create challenges and we are concerned about the possible risks and consequences relating to working conditions, skills and training, ethics, equality, health and safety (among others). Therefore, CFDT Cadres would like to underline the importance of addressing AI technologies and robotization as topics for collective bargaining at all levels (company, national and European). AI and robotics have a huge impact on the future labour market, as jobs will sometimes disappear or be transformed and other jobs will be created. We need to accompany this process and address the question of skills and training for the future workforce: need to ensure that training on necessary digital skills is provided by education institutions and companies, and that it is not the sole responsibility of the worker to keep up with the rapid technological developments. Employability needs to be promoted through upskilling and reskilling schemes for workers. Investment in formal, informal and life-long learning is key; we must enable people to work with AI or invest in competences that AI will not cover. It is important to develop action plans at EU and national level together with education providers and social partners in order to modernize education and vocational training. We therefore welcome the call from the ILO Global Commission on the Future of Work for "a formal recognition of a universal entitlement to lifelong learning and the establishment of an effective lifelong learning system. (https://www.ilo.org/global/topics/future-of-work/publications/WCMS_662410/lang--en/index.htm)- The social partners play a key role in this and the EU should cooperate with them and national governments in order to identify which job sectors will be affected by AI. We need to understand the timeline and extent of changes in the labour market. The involvement of social partners is a must to find appropriate and future-proof solutions to concerns relating to employment, training, the nature of work, (in)equality or social systems and collective bargaining, especially at sectoral level.- As AI and automation have the potential to transform not only simple tasks but very complex processes, we need to have a large public discussion about the areas in which the use of AI is reasonable and beneficial for society. Part of the debate should be the question of how the profits generated by AI should be re-invested for the common good by creating employment in domains such as care, health services, education or mobility. Employees should participate in the distribution of profits, e.g. through wage increases or reduction of working time. Moreover, AI wins should be used to strengthen social security systems. This could be a measure to address the problem of future job losses and the precarisation of employment relations in a-typical work (e.g. platform work) due to AI and automation.- It	- CFDT cadres supports the human-centric or human-in-command approach suggested in the guidelines. We agree that it is necessary that humans always need to remain in control of technology and machines. Likewise, we agree that the use of AI needs to respect European values and fundamental rights.- We recommend an expansion of 3.2 "In an AI context, freedom of the individual requires protection from direct or indirect coercion, surveillance, deception or manipulation." Much profiling that lies at the heart of AI systems relies on a degree of manipulation of data. This is not least relevant in relation to the world of work and especially in the use of AI in Human Resources. - We welcome that the HLEG understands the need to ensure that those involved in the development and marketing of AI (researchers, engineers, designers etc.) act in accordance with ethic and social responsibility criteria. This should be addressed by changing educational priorities for technical subjects and by providing lifelong learning opportunities (e.g. by incorporating ethics and the humanities into training courses in engineering). - Organisations and companies should develop tools to facilitate ethical discussions and decision-making throughout the whole design process. This should be completed by internal training programs on ethics for all employees. Such training should help employees understand the AI systems themselves, their rights in relation to said systems and their possibilities of redress, complaint etc. - AI should provide an opportunity for workers to apply their skills and competences to the fullest while at the same time remain owner of the production process. This includes the principle of transparency in the use of AI systems in HR, like the hiring of employees or the performance assessment of staff. It is important to safeguard the rights and freedoms of employees in line with non-discrimination rules as regards the processing of workers' data.- CFDT Cadres welcomes 5.1 – 5.4. We support that these examples raise real-life concerns of the adverse consequences of AI systems. - In 5.2. CFDT Cadres urges the group to expand on the issue of the human's right to know they are interacting with an AI identify. This could be done through a "labelling" system. For example, online bots should be labelled as such. Users should be made aware of the use of bots and AI in customer call-centre or help desks etc. - We would welcome that the employer-employee, employer-worker relation is explicitly mentioned in 5.3 as an example of power asymmetry. - Taking into account the power asymmetry in employer-worker relations, a separate point 3.6 on "workers' rights" should be added, which should contain the following points: "decent work by design", equal negotiation processes in the sense codetermination rights, informational self-determination of employees, non-discrimination principle and freedom of association including the right to strike. This is needed in order to secure worker's rights to co-decide on aims and application of AI systems, and create a legal framework.- Concerning the long-term risks and concerns we welcome that these should be considered. This could become an integral part of the accountability and audibility demands – i.e. that developers, users deployers etc need to reflect on the	- In order to achieve "trustworthy AI", we need to establish public, independent and autonomous organisations that can control and audit (labour) algorithms (e.g. to identify underlying biases and the objectivity of data sets that train algorithms). Likewise, the implementation of the ethical guidelines on AI must be monitored. A European observatory focusing on the ethics in AI systems could play the role of an independent watchdog, including in business.- We would like the advice, to always keep record of the data that is fed to the AI systems" from the heading of data governance included under Accountability. For workers, it is paramount that the datasets used to evaluate performance, or in hiring or firing processes is transparent and can be accounted for. - In companies, managers are concerned first and foremost in digital transitions. They have an essential role in managing these changes and introducing technologies. They can guide them and propose solutions to the dilemmas they can always generate, particularly in the field of recruitment and HR. This requires training and real flexibility for managers.- The explanation of the principle of autonomy covers the question of AI at work only in a footnote, whereas this is an important issue that should be given a more prominent place. We would like to highlight the right of workers to individually and collectively opt out or withdraw from the use of AI systems (or a decision chosen by an AI system) if they undermine the workers' autonomy, decision making competence or disrespect fundamental rights and ethical principles. We recommend the inclusion of a special chapter that provides for ethical guidelines on AI in the work environment to address these issues more in detail.- CFDT Cadres welcomes that the HLEG on AI acknowledges the importance of social dialogue to realise trustworthy AI. We would like to add that the involvement of social partners, and in particular employee representatives, should not only take place regarding the general public debate on AI. Social partners should be involved in the establishment of codes of conducts, of standardisation schemes, development of training and in the proposed accountability governance. Employee participation and inclusion should take place early in the design, development and deployment of new technologies including AI and robotics. It is essential and important not only to inform and consult workers representatives in the work place or at branch level, but to enhance their co-determination rights and ensure their right to co-decide on the aims, reasons and implementation of AI at the workplace. - Social partners at all levels should be involved in the implementation at company, industry, national or international level, including through collective agreements setting standards. In this context, it is required to describe the negotiation processes, e.g. central control structures for sector-specific solutions (cf. ‚AI Now' Report 2018): „Governments need to regulate AI by expanding the powers of sector-specific agencies to oversee, audit, and monitor these technologies by domain." (https://ainowinstitute.org/AI_Now_2018_Report.pdf)- Regarding the principles of accountability and transparency, we need to establish mechanisms for the protection of whistle-blowers who disclose the risks of AI	- We welcome that processes shall be examined in order "to allow a human control, if needed" (assessment list – governing AI autonomy). In so doing it should not be a question "to keep a human in the loop". We need clearly defined measures, that empower people to exert this control in all processes – regarding resources (technical equipment etc), organisational needs (time, liability, etc.) and qualification.- We suggest extending the list on the assessment of use cases (p.28) and add the question of processes, in order to use AI to ensure decent work (development and impact assessment).- Concerning the question in the final note p. 28: In the case of the specific field of healthcare, the medical secret should not be neglected and sensitive information should in no case be transmitted to the insurance industry for the setup of fees.	- CFDT Cadres welcomes the call for Accountability Governance on page 21. The establishment of Data/AI Governance Councils in companies will indeed strengthen the accountability of AI systems and will address a weakness in the GDPR. The Council should consist of shop stewards and management and be responsible for holding management accountable and transparent to the use of AI and data. Whistleblowers should be able to address concerns to the Council and mandate the council to investigate on reported issues. - We welcome the process of developing guidelines for a trustworthy AI made in Europe, which encompasses corresponding "guidelines made in Europe", but would like to raise the question why non-European companies such as Google were granted full membership and full participatory right in the High- Level Expert Group. The status of associate expert would be more appropriate.- CFDT Cadres also supports the position of the ETUC regarding this consultation.

				is therefore important to integrate the aspect of the quality of jobs, decent work and social progress into the ethical approach in order to create a balance with the purely economic objectives underlying the creation and use of AI and robotics. - The Human-centric approach (HCD) not only presupposes information, transparency, participation and traceability, but also requires specific negotiation processes regarding decision-making in view of the aims and implementation of AI-systems at a very early stage for stakeholders such as employees and their codetermination bodies.	development/changing nature of the adopted AI as well as engage in predictions/forecasts of its future development scope and the consequences (positive and negative) hereof). - AI's influence does not only affect the world of work, but also democracy and society as a whole. We welcome that the draft refers to this point in Chapter I, paragraph 5.3., by stating that AI is not to be implemented in order to enable "citizen scoring" by a state/government. But this should also apply to private businesses. Neither states nor companies should be allowed nor have the possibility, to create human profiles such as "moral personality" or "ethical integrity". We reject the proposed opt-out-function and even possible "opt-in"-functions are not to be designed in a way that they conflict with fundamental human rights and possibly lead to the waiving of services that are useful for a person. AI-based services, that are important for work and life, must be designed in such a way that they do not require the collection of data which could be useable for human profiling.- Creating big data-bases always includes the risk of hackability as well as intentional and unintentional data-leaks. The guiding principle of "data-sovereignty" needs data-security in order to be viable. This implies explicitly not surveying data in areas that are of highly explosive nature for people in e.g. political, private or work-related areas. Fundamental rights as informational self-determination, the freedom of association and freedom of speech are not to be put at risk by creating such data-bases.- Freedom to opt out:- In addition, the freedom of an individual may be seriously hampered by the use of AI: for example, if a person is obliged to provide medical data through connected objects to get insurance. A right to opt out, without fear of retaliation or any negative impact (for example a forfeit payment when not delivering data from applications) is therefore an absolute necessity. The right to opt out is explicitly stated for citizens and government action in relation to citizens (under paragraph 3.5 and 5.3.) We believe that this right should also be extended to consumers (i.e. for insurance, credit obtention, etc) and to workers.	systems or the non-respect of ethical principles – especially in the case of employees in companies that develop such systems. Internal reporting of risks and violations should be supported and rules in place to ensure follow up. - Organisations and companies should pay attention to potential biases encoded in the system development, training data and model performance – especially those that my affect the most vulnerable. They could also establish an internal ethical review process to democratise the decision-making process- Companies should not only increase transparency regarding the design and development of AI systems, but also in organisational chains of responsibility. - CFDT Cadres calls for European ethics committees and in each member state, the transparency of algorithms to try to control potential abuses (related to gender, culture, objectives... of developers and their organizations (see the Facebook algorithm which overvalues discussion groups vs. the pages of organizations). - These ethics committees should have the power to impose sanctions, like the CNIL in France.		
Anonymous	Anonymous	Anonymous	None.	Critical concerns raised by AI: Potential longer-term concerns Today a large number of AI systems are being used on our daily lives. Humans and AI are already closely intertwined and odds are the connection will only get closer. AI will keep growing and get to be integral part of human societies, and so it should be built around ethical foundations and values. Even though current AI can be considered narrow, systems are getting better every day at their tasks, pushing further the concept of what a narrow AI can do. One milestone after another has been reached. Both futurists and scientists have analyzed the progress of technology and determined the growth is exponential, which makes it really difficult to make predictions. We would likely experience several years of progress in 2019 if considered at past years' rate, and that could be the road to a broader AI. We should not ignore the risks involved, even if they might now seem like science fiction, as the exponential trend may get	Technical and Non-Technical Methods to achieve Trustworthy AI: Technical Methods: Traceability & Auditability, Explanation (XAI research) There already exist interpretability tools which help humans gain some understanding in the way the algorithms behave. These tools might not be able to capture the behaviour of the model as a whole but rather provide explanations on how particular decisions came about, a suitable approach even when working with complex neural networks. These tools can assist in traceability and auditability tasks, and incorporating them into the development process of AI systems can help engineers identify risks, biases or errors, or for example verify if a system's good performance is not due to undesired patterns in the training and testing data.	Assessing Trustworthy AI: Transparency: Traceability. Method of testing the algorithmic system: ethical-dilemma tests would be mandatory Outcomes of the algorithmic system. Not only the outcome should be provided, but a brief and comprehensible by human beings explanation would be needed in order to ensure transparency and earn trust in AI systems.	As discipline, the proposed definition for AI should be improved. Robotics is not a discipline of AI. Robotics may make use of AI, for instance, for cognitive functions which not always applies to robotic systems. Responsible research, responsible innovation and responsible business must be in the scope of these guidelines. A taxonomy of IA applications together with potential impact and side effects would be needed. As a consequence, ethical-dilemma tests would be mandatory.

				them closer than we assume based on the growth rate so far. Humans' relation with AI is and will be very deep, which leads to huge risks. A risk-assessment approach seems therefore the wisest.			
Anonymous	Anonymous	Anonymous	Test To Ignore	Test To Ignore	Test To Ignore	Test To Ignore	Test To Ignore
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Anonymous	Anonymous	Anonymous	At first, we support the need for a ''human-centric approach'' to AI as well as the two components of trustworthy AI: to ensure its ethical purpose and its technically robust development. In this sense, we think that the real added value of this document is precisely the guidance on the concrete implementation of ethical principles into AI systems, provided that we can indeed achieve in create such implementable guidance. Furthermore, we welcome the intention to engage in a debate at a global level. For this reason, we support heartily the ethical approach to AI as a key to enable responsible competitiveness, because the international competitiveness will not just be dictated by whom spends more money but, more importantly, by whom develops the right AI. That's critical for Europe. In other aspects, though, we think that bias is not always bad, but it could be intended and positive sometimes. Even so, we need to make sure bias does not lead to unfair discrimination. On the other hand, we share the same opinion on presenting the different components of trust, including the technology itself, its rules, laws, norms and public governance models as well as the business developing. Despite this document should not be seen as an end point, we need to also make the point that this document should be used as a reference in national discussions on AI (especially for the creation of national AI plans). That would be the sense of a being a ''beginning of a new and open-ended process of discussion'' because it could set the rules of a fair discussion about IA. We also need some more information on the process through which the document will be updated after its publication. We believe that it should continue through the HLEG, but need a regular engagement with other stakeholders and a formal consultation process. Finally, we share effusively the need of a tailored approach to each situation, given AI's context specificities, because each one raises different challenges.	Of course, we support the ethical purpose on the approach to trustworthy IA, as stated earlier, but we have some questions regarding the point Fundamental Rights of Human Being: What happens if there is a conflict between rights? Is there a hierarchy that needs to be respected? As the draft correctly points out, ''in particular situations, tensions may arise between the principles when considered from the point of view of the individual compared with the point of view of society, and vice versa''. We see the need to elaborate on this because developers and companies need more guidance. The really need to understand if there is maybe a hierarchy that needs to be respected. Someone has to decide what to do when there is a conflict between rights. On the other hand, regarding the Freedom of the individual right: How does it fit with national security obligations and requirements of a government? Also, doesn't one freedom stop when it starts infringing on the freedom of someone else? However, we support companies having AI Ethics Boards, as well as the right to a human-centric appeal of decisions made by AI systems. We share the work of AI4People's project which is used as a basis for the five principles too. Regarding concretely to the Principle of Autonomy, we support individuals having the right to know if they interact with AI or not as well as technological and business transparency models, matter from an ethical standpoint. However, we consider that the development of new means by which citizens can give verified consent to being automatically identified by AI could be problematic since it will block innovation. With regard to GPDR, we think that the usage of ''anonymous'' personal data that can be re-personalized seems to go beyond GPDR guidelines. Also, we have to add that under GPDR there are also another legal basis such as legitimate interest, not only the achieve of informed consent.	The Requirements of trustworthy AI list seems too long. We suggest merging some items and focusing more. A possible solution would be to merge Data Governance and Respect for Privacy; Design for All and Non-Discrimination; Respect Human Autonomy and Governance AI Autonomy and Robustness and Safety. Returning to the bias in the field of Data Governance, it needs to be clarified that some kind of bias is good unless is intended because of the objective for the AI system. It seems necessary to clarify that the design of the systems ''in a way that allows all citizens to use the product or services, regardless of their age, disability status or social status'' (at the third point, Design for all) is related to ''Accessibility''. On the other hand, with regard to the Robustness, we think that someone, somehow, needs to decide what level of accuracy is acceptable for an AI system in a certain use case. Furthermore, the fallback plan seems to be an interesting idea which needs to be further explored if feasible. However, it also depends on the use case. We also support effusively the transparency point (in special the idea of being explicit and open about choices and decisions concerning data sources, etc.) as well as the prominence of evaluation and justification processes to the development process. Finally, regarding the Non-Technical Methods, we think that the standardization needs to be linked to ''ethics by design''.	No further comments	No further comments

Anonymous	Anonymous	Anonymous	• Adding some major academic references on “positives and negatives” – page 1 - (particularly on the fact that “on the whole, AI’s benefits outweigh its risks” – page i - would constitute a good introduction and bring a strong foundation to the rest of the document.	• Very well documented chapter, particularly with the list of existing principles taken into account and analysed (section 4). • Pages 8-9: The principles of beneficence (“do good”) and of non-maleficence (“do not harm”) are very close, because they are counterparts. The purposes of “generating prosperity, value creation and wealth maximization and sustainability” (mentioned in the “beneficence” section) and of “protecting the dignity, integrity, liberty, privacy, safety, and security of human beings” (mentioned in the “non-maleficence” section) are all objectives of “doing good” and “not harming” after all. Both principles and associated examples should be clarified. • Page 11: “Identification without consent”: Giving examples for the cases “Where the application of such technologies is not clearly warranted by existing law or the protection of core values” would enable an easier understanding of this section. • Page 12: “Lethal Autonomous Weapon Systems ». “Ultimately, human beings are, and must remain, responsible and accountable for all casualties.” This sentence related to the decision power of machines not only applies to military issues, but also to civil systems: such as automated cars and the potential decision to be taken by the machine – when an unexpected event happens - between impacting pedestrian 1 and impacting pedestrian 2. So section 5.4 could be larger than LAWS-focused and incorporate all types of machines decision-making.	• Page 21: “Non-technical methods” could be classified in order of importance by the AI HLEG and this should be mentioned in introduction of the section. In my opinion, education and diversity are the 2 most important methods listed here. • Page 21: The issue of customer data ownership in private companies (ownership by the company vs. the customer) could be part of the discussion in the “regulation” methods. • Page 21: “Interoperability” could be added to “standardization” in order to contribute to transparent and open AI systems, and thus to trust (Collaboration of governments and industries is key in this area). • Page 22: “Education and awareness to foster an ethical mindset”: This section should insist more on the fact that awareness is a way to empower all citizens by enabling them to master and understand the data and technology. Education at an early age on these subjects (demystifying technology and AI) is key to the informed consent of citizens and a way towards improved AI ethics. • Page 22: “Diversity and inclusiveness” not only apply to the AI design teams, but also to the data sets collected and analyzed by the AI.		• The report looks very well designed. It probably lacks some illustrations and examples to be easily readable by people who are unfamiliar with AI issues. • Another interesting issue to raise regarding AI and ethics is more market-related: the asymmetry of information between on one side a small number of big companies who own/control big data and on the other side, small companies.
stefan	koreneef	NL Ministry of Economic Affairs & Climate, Ministry of the Interior & Kingdom relations	Support: NL supports the initiative of the European Commission and work of the High-level Expert Group on Artificial Intelligence (HLEG AI) which is a good -diverse-representation of businesses, NGO’s and social partners. The draft guidelines strike a good balance between risks and opportunities, AI should be trustworthy and respecting fundamental human rights in applicable, responsible manner. From the Member States’ perspective these guidelines should be an element of the different national AI strategies. Which, in turn, are in line with the EU Coordinated action plan.• Tone of the document: we should have more emphasis on how to stimulate good AI, rather than control bad AI (now only the conclusion contains a more positive tone). We want to achieve trustworthy AI, that will be seizing the opportunities and creating a leading position for Europe on AI.Rationale and foresight of the guidelines, five suggestions:o The goal and purpose of the draft guidelines are not completely clear. How can we use the guidelines? If it’s not going to be a directive, what can we do with it? Suggestion 1: we should add more European global thought leadership to the guidelines and ambition: a statement that the EU wants to become AI leader. Europe will only be able to set standards if it has something to say in terms of innovation and adoption and then it is more logical to incorporate it in all national AI strategies of Member States. The global reaction to the EU’s General Data Protection Regulation (GDPR) has shown a strong framework which can shape global markets and strengthen the EU economy. o If the EU wants to be a leader, it is good to have our own principles. A lot of the principles based	Ethical principles (page 8 to 10):• Ethics by design should – from a technical perspective- be further examined with ethics by the adaptive system and ethics by behaviour.• The term ‘explicability’ is problematic (if required in common language), explainability could be an alternative We would suggest to highlight the 3 terms: describe, inspect and reproduce (= which de-facto translates into auditability)• Related to this, some concepts seem to be used interchangeably: explicability (p10, but is not used in the rest of the doc), traceability (p20, which might mean the same, but only in a specific case), explainability (p21) are all used, but seem to cover more or less the same idea. We would suggest to Please clarify these concepts and reduce the use of different concepts throughout the document.• Problematic: training data also falls under GDPR (for example: how do we define consent for training data) Is there a way to exempt specific consent requirements (for example in public health sector)• Standards: EU first then adoption elsewhere, or global from the start• Accountability governance: suggested title for function in organizations: Data stewards? (already used for data analytics governance, and gained traction recently. Building on what is accepted might help, rather than coming up with new ideas. • Human oversight: state the need to add resources to organize frequent conferences for (international) knowledge exchanges – EU must frequently showcase what is being taught/applied.• R&D: traceability and explainability: how decisions come about (not why) – need for more research (XAI): state that EU should work on incentives from public agencies to stimulate this type of R&D	Specific points, in addition to the general impression• Page iv glossary, definition of AI: systems that act in the physical or digital world, by perceiving, interpreting, reasoning and deciding.” This definition assigns technology in a way that could be misleading: nobody talks about a thermometer “perceiving” temperature, although it “decides” very clever “how many degrees it is” based on “what it perceives”. Just like a thermometer an autonomous car, robot or character recognition device does not perceive anything. The machines are configured (not “trained”) to map input on a predefined output: turn the steering wheel, initiate or stop a process or output something. So, when an AI device “does not understand me”, this means it cannot match the input with an output. This happens either because the input was never presented before or because there is not a suitable output.• Page 1, Trustworthy AI, a key element. We would suggest to add numbers: how many people want trustworthy AI and an increase in research on the importance of trust. • Page 2, role of AI Ethics. It is a starting point, not the finish line. In this section the purpose and role of the guidance can be added. Also the importance of experiments as well as trial and error. It could be made more vivid by turning it into a tool, using best practices and results of previous cases. • Page 3, Scope of the Guidelines: if you describe the scope, describe what it is rather than what it is not.	Page 3, A Framework for Trustworthy AI. realisation of trustworthy AI. We would suggest to include human rights and make our own (European) principles. Use cases: make it dynamic, trial and error and lessons learned.	Greater emphasis for selected issues: • Validation is important: we should suggest to further define what is tested and what is not, as well as how the testing will be done.• Good graphics: the greater the feedback loops, the fewer testing requirements• The four user cases for corporate and public AI mentioned in the document should be presented at the beginning of the document (it’s a good delivery of the HLEG AI and follow up). • The tensions concerning transparency, i.e. between transparency and innovation (when making public business secrets) and transparency/explicability and added value of AI (when demanding transparency or explainability means that specific types of AI, like deep learning neural networks, cannot be used as they are inherently opaque). This is very briefly addressed in the summary and on page 23 (i.e. gaming the system), but is not further illustrated. Specific additions: • Failsafe shutdown is a good proposal; but we might need a European or international arbiter to keep record of when it was used, by whom. (this could be an additional item for non-technical methods section)• Algorithms: not in the glossary or specifically mentioned, we would suggest an extra paragraph • Transparency: add the question on whether a warning sign for user might be useful whenever a personal ID is determined / used? Think of incentives not to go all the way at once, but to start with supervised AI (ethically easier stuff)• “Common good” is very promising, but hardly elaborated: common good is interpreted (as elaborated in the report) in a narrow way as “contributing to a good life”. That is a reduction and potentially harmful to ethical application of AI. Therefore “common good”

			in other continents are developed in, for example, the BIO tech industry (page 6 of the draft guidelines). o Suggestion 2: We should be focused on our own European principles, including human rights and a human centric approach as mentioned in the guidance. o In order to be leading, the EU must find a way to approach AI from a global perspective. We cannot exclude products and services from outside Europe. What are we going to do with AI products in the EU that would not be compatible with our European framework, and AI products that may be applied in line with our framework, but that are trained or developed in a way that are not compatibleo Suggestion 3: Make the guidelines more globally oriented: it is therefore important to work together as Europe in other global fora, for example the OECD, ITU and G20.o The roles of private industry and public sector need to be further determined. How do we make the guidelines operational? o Suggestion 4: It can be made practical through experiments. Trial and error, ethics in, by and for design is creating a new situation with trustworthy AI. It also tackles the dynamic and pragmatic reality. Standards and self-evaluation can play an important role self-regulation will only work if we have smart evaluation or ISO/audit systems to make this work. That will speed up our own learning curve on trustworthy AI and foster the willingness of consumers and institutions to adopt AI-applications, which could offer opportunities for Europe. o We would suggest to refer to existing European Commission Better Regulation documents or OECD documents, like the OECD Due Diligence Guidance for responsible Business Conduct.o Suggestion 5: Three elements that could be used as the key organizing principle for the public and private sector: ethical purpose of the guidelines, technically robust guidelines and a leading role for Europe at a global stage, this could be reflected in the title.	so that EU becomes leader in XAI as a growing academic field.• New categories: when working on possible codes of conduct: introduce specific types of AI (for example: decision-support is different from autonomous systems, lethal force is different from consumer products)• Suggestion to the European Commission to establish EU and/or national awards for best practices from private and public entities.Suggestion to establish academic programmes to work on trustworthy AI (MA programmes for example) in close cooperation with the Member States, academia and businesses or, suggest a EU Erasmus type programme to help stimulate uptake across Europe		by “well-being” as defined by the OECD• Another non-technical method is the implementation of an AI Impact Assessment. In the NL we have published an AI Impact Assessment”. This could be designed in line with the design of a privacy, data protection or human rights Impact Assessment. A strength of this approach is that it can properly take into account the specific context.
Anonymous	Anonymous	Anonymous	Concerning 5 principles adopted by the draft document (Do good; Do no harm; Preserve Human Agency; Be Fair and Operate transparently) which are generally good – meaning, given the national, cultural, religious and worldview differences of individual Member States, these principles are broad enough to be adopted by all. Nevertheless, these 5 principles have to be specified considering particular national (legislative and social) specifics of the Member State. However, there are some important things missing here. For example, we can ask if it is ethical to apply AI to increase profits on the account of employment or the role of people? Therefore, I suggest the addition as following: When defining the main principles (Do good; Do no harm…) in the Chapter 1, section 4, we should mark that these principles depend on various national, cultural, religious and worldview differences of individual Member States. Ad Chapter 1, section 5 - Privacy. Privacy is insufficiently mentioned through GDPR, what is very important issue because the perception of privacy through the development of digital technologies has been redefined several times. In general, the ethical principles that are mentioned do not	Ad Chapter 2, section 2. Technical methods. Ad. Requirements of Trustworthy AI which are well-defined, but there is lack of measurability. Measurability parameters are very important because we must be able to measure what the AI system does. So I suggest that the AI system manufacturers already in this early phase try to elaborate certain measurability parameters and this should be added in the Chapter 2, section 2 as Measuring& Testing&Validating. Ad Chapter 2, section 1, subsection 4: I think that in this chapter the definition of the security levels that monitor the intrinsic and extrinsic behavior of the system should be more detailed. Ad Chapter 2, section 2, subsection Traceability & Auditability: I suggest the sentence ''Each AI system must transparently declare the sources of the data he's using.'' to be added somewhere in the text. Ad Chapter 2: Since this type of problems isn't particularly addressed anywhere in the draft, we would like to draw the attention to the following: *How should we embed AI systems in our social relations? *How we will address the problem of unemployment if AI makes lots of workers	Cognitive bias causes systematic deviation from the norms or rationality in judgment. All available data are always limited, and as such, they are assumptions of deviation / bias. That is why cognitive bias is one of the biggest problems of modern AI. For example, bias can cause unintentional damage if the designer made a mistake during embedding an algorithm into the system. This is why I suggest the introduction of measurable parameters for bias in chapter 3, subsection 4 Governing AI autonomy. Which control mechanisms should we use to prevent such issues to happen? Given the fact that AI is not just a short-term trend, but a system that defines the future of every successful business, companies are under great pressure trying to balance the use of artificial intelligence to improve their business strategies while simultaneously protecting user privacy. So I suggest an addition to the assessment list concerning subsection 4 Non-discrimination: How to balance the relation between the profit and potential security flaws?	I am thrilled that so much attention is devoted to the making of ethical framework for the implementation of AI. After all, since artificial agents are getting smarter and smarter, it is in our best interest to ensure safe and successful application of robotic technology. However, there are few things I would like to draw Your attention to: 1.The definition of AI: The given definition includes several disputable terms: 1.Reasoning. We should make a brief distinction between systems that act and systems that reason. Acting like someone/something doesn't necessarily include any type of awareness of actions, but reasoning is very powerful cognitive action which is attributed exclusively to humans. Being able to reason means some much more than being able to be logically correct and classify data. So, if we take into account the specificity of artificial intelligence as technology, the expression ''reasoning on the knowledge'' is, scientifically speaking, extremely inaccurate. Since the improper or vaguely use of terms has very important philosophical consequences, in order to avoid potential misconceptions, I suggest a brief definition to understand exactly what is

				represent anything new and can be applied to any technology and do not say anything about the specificity of AI. It is known that autonomous systems independently shape their behavior. How to deal with this issue? Such systems can shape some form of ethics as well, which we must be able to recognize and direct (which is opposed to "technically robust" what is a prominent term in the draft). Ad Chapter 1, section 5.4 - LAWS - Military purposes. In relation with the use of artificial intelligence in military purposes, the fact that needs to be not just clearly emphasized, but prohibited by the law is the uncontrollable use of AI for military purposes against people.	redundant? Although it is unlikely that AI system could ever reach such a high level of autonomy in order to operate completely without any kind of human control. This is why I suggest these issues to be examined and added to the Requirement list in the Chapter 2, section 1 as 11.social impact.		meant by each term used in the definition. Since there already stands the expression ''according to pre-defined parameters'', I propose changes as following: First ''according to pre-defined parameters made by designers'' and second, instead of the term ''reasoning'' I suggest the term ''interpreting''. 2.Deciding the best action. The greatest ethical concern arises when it comes to deciding the best action - Does decision making mean that AI must be a moral agent as well as human person is? But how could AI be a moral agent? To be a moral agent, X must satisfy certain criteria such as 1. Consciousness, 2. Self-consciousness 3. Free will, 4. Autonomy, 4. Imagination and so on. Considering that AI is understood as intelligent mechanical system which measures itself by how successful can be copying and interpreting human ways of behavior and mental powers, it is pivotal to provide a moral basis for those actions. AI consists of algorithms designed to solve a variety of problems. But algorithms are not objective truths - humans who create them, embed their values into algorithms. Machines don't have goals or desires of their own – this is why the responsibility for the actions of AI lies with their manufacturers, retailers and users. Since ethical values are inherently human, the question is what kind of ethics or moral codes should we implant into the AI system and what follows from a given moral code? Robotics becomes ubiquitous in today's society, so the need for answers concerning this specific domain will certainly going to grow in importance and extent. Therefore, besides the expression ''deciding the best action'', I propose the addition as following: ''deciding the best action (''according to pre-defined parameters made by designers based on specific ethical norms/values'').
Danny	Van Roijen	COCIR		Right to decide to be subject to AI/right to opt-out/ right of withdrawalPages 10 - 11 include a right to either be subject to AI/a right to opt out and/or a right of withdrawal.A right to decide to be subject (or not) to AI, a right to opt out and a right to withdraw significantly reduces the possibility to make use of AI systems. By definition, AI relies on large volumes of retrospective data, making the execution of these rights impossible for any AI system, especially since typically AI systems will further use the input by users to improve the algorithms the AI system is built of.In addition, these requirements were not omitted in the GDPR. On the contrary, GDPR, which regulates data protection, provides already for very specific requirements regarding automated decision-making. As an example, page 10 stipulates:"If one is a consumer or user of an AI system this entails a right to decide to be subject to direct or indirect AI decision making, a right to knowledge of direct or indirect interaction with AI systems, a right to opt out and a right of withdrawal".We propose to limit this sentence to the following:"If one is a consumer or user of an AI system, this entails a right – at any time during the use – to decide to be subject to direct or indirect automated decision	Page 15 includes an unclear statement on data governance which should be deleted and it is: "To trust the data gathering process, it must be ensured that such data will not be used against the individuals who provided the data." The AI guidelines should not lead to a situation where for instance patients whose personal data was used for development of an AI system that can detect cancer cells, cannot profit from the future use of that AI system to have their own cells checked. Page 15 also includes a requirement "design for all". This paragraph needs to be clarified, to reflect that AI systems can be designed for specific user groups and need to be (only) user centric for the targeted user group. For instance, AI systems intended for use by medical specialists do not need to be tailored to the lay knowledge of the average individual, as the systems will only be used by medical specialists.	All the comments made above also apply to Chapter III.More in particular, the above comments should be specifically reflected in the following items: • 3. "Design for all":o "Is the system equitable in use?o "Does the system accommodate a wide range of individual preferences and abilities? "• 6. " Respect for Privacy" o "How can users seek information about valid consent and how can such consent be revoked"	

				making, a right to knowledge of direct or indirect interaction with AI systems."In addition, page 12, paragraph 5.3 includes a right to opt-out from any scoring mechanism: "and ideally providing them with the possibility to opt-out of the scoring mechanism" and also states: "Developers and deployers should therefore ensure such opt-out option of the technology's design, and make the necessary resources available for this purpose." It is difficult to see how to comply with such a requirement, as data will be interwoven with the algorithm. Therefore we propose that this is limited to situations in which consent is the legal basis for the personal data processing by the AI system. If not consent, but another legal basis is used, for instance legitimate interest, there should be no requirement with regards to opt-out functions.Informed consentPages 10 and 11 refer to informed consent. It is not clear whether the document prescribes informed consent as a hard requirement for any AI system. The assessment in Chapter III (also see below) seems to indicate that this is the case. Basing the data processing for AI exclusively on informed consent will seriously hamper the use of AI, as it leverages large volumes of retrospective data. We believe that the GDPR safeguards provide sufficient protection for any AI system, as AI is a specific form of data processing. GDPR already ensures a legal basis, transparency, explicability, human intervention in automated decision making and accountability. However, GDPR identifies six legal bases for processing personal data, of which consent is only one.		
Anonymous	Anonymous	Anonymous	- The document is stated as aiming the development, deployment and use of AI. I am wondering if it is a two broad audience to aim for, thus leading to a very high level document that feels more like wishes that actual guidelines. - The guidelines should be written in a way to lead to actual regulation instead of a voluntary endorsement.	- transhumanism should likely be mentioned as it can fall under AI in some cases and could have huge impacts - the right to opt-out is a real challenge. if deciding that we don't want to use AI means that we should chose an other service provider. And it should be coupled with incentives for businesses to offer a modular service for which people still get access to it (or a meaningful portion of it) even without giving personal information.	- The listed requirements could apply to almost any technology. How are they specific to AI ?		- I didn't see anything about the impacts on society, wealth gap, social behaviour, ... This is lacking. - The definition given for AI as "built by humans" is wrong according to me. autoML is not actualy built by humans.
Julia	Wadoux	AGE Platform Europe	AGE fully support the rationale and foresight of the guidelines in particular the human centric approach and the respect of fundamental rights as the basis for its development, whatever the tailored approach that may be used at a later stage to answer specific issues related to a domain. While it makes all sense to consider that the document is the starting point for debate, it should also be noted that if the aim is to have formal endorsement of the ethical guidelines by stakeholders, there is a need at some point to have a final version or at least to find a way to renew endorsement if the guidelines are amended. In addition, it should be clear for stakeholders endorsing the guidelines what has been amended for them to see whether or not they can continue to endorse the guidelines. As human rights are universally accepted and do not need to be revised, ethics guidelines for AI should at some point be	Since it is referred to the Oviedo Convention on page 6, it is important for us to underline a concern AGE has in relation to an additional protocol to the Convention which is under negotiations. This protocol refers to the "protection of human rights and dignity of persons with mental disorder and regard to involuntary placement and involuntary treatment". We have raised concern on the draft additional protocol since we consider it runs against existing human rights standards, give extensive power to medical professionals and breach personal dignity and autonomy, which should be guaranteed regardless of age or disability. Therefore, we would like to underline that the AI HLEG is using the same kind of process as the Oviedo Convention which use fundamental rights to derive ethical principles and values, we shall be clear that this requires a comprehensive approach across the process and not create exception which would dilute the fundamental rights approach.	We very much agree on the list of 10 requirements. At the cross-roads between Data Governance and Non-discrimination, we would like to highlight how much important it is to improve the set of data in order to reflect the population composition in the best possible way. For instance, even if older persons are less using new technologies and mobile applications, they should be reflected in the way systems are developed considering they represent a high proportion of the population. Likewise, the gender dimension is critical (and even more among older persons). This is particularly true when it comes to AI in the healthcare area since older persons are the vast majority of the care recipients. Regarding the part related to technical methods to ensure trustworthy AI, it could be meaningful to better echo requirement #3, i.e. Design for all which is very much linked to the involvement of users in the design and development of solutions (user involvement, co-creation). Indeed, while it is important to include a wide range of user	In relation to the point raised above in relation to user involvement and co-creation, it might be useful to add a point under #3 Design for all: - How users were involved in the development of the solutions? - What kind of users have been involved? And why? As for the four areas which have been chosen, here are some comments around the first three: (1) Healthcare Diagnose and Treatment: - the quality of the dataset is really critical and must reflect the population in the best possible way, in particular older persons and women. This is an issue which is already at stake today when it comes to clinical trials and medicines, so that it would be a pity to repeat the same situation. - It would be important to consider that you have mainly three different types of end-users: medical/care staff, patients and their informal carers. The three are important and	All in all, AGE Platform Europe welcomes this work conducted by the High-Level Group on AI and very much support the approach based on fundamental rights' approach and will monitor the next steps with great interest.

comprehensive enough to be valid now and tomorrow.
Eventually as it is the case for human rights, there might be a need for a more detailed and clarified piece of work for specific situations (e.g. UN Convention on the Rights of Persons with Disabilities do not create new rights but basically explain how Human Rights should be applied in the context of disability). But these specificities will be covered thought the assessment list (Chapter III) so that the ethical guidelines could be a stand-alone document which wouldn't need to be revised every year.

We very much support the whole chapter 1 and would like to provide with an additional source of information. In July 2017, the UN Independent Expert on the enjoyment of all human rights by Older persons has released a report on the impact of assistive and robotics technology, artificial intelligence and automation on the human rights of older persons. This is a very valuable piece of work which provides with meaningful and balanced insight on these issues.

The report is available at: https://www.age-platform.eu/sites/default/files/Report%20of%20the%20UN%20Independent%20Expert%20on%20digitalisation%20and%20use%20of%20robots_2017.pdf
In this report the issues raised under point 5.1 "Identification without consent" is also addressed, notably when it comes to the GPS bracelet used for people with dementia. These bracelets offer an interesting opportunity to physical or chemical constraints, but their use should also be part of a larger conversation on informed consent and transparency when it comes to the data used or retrieved through such tools.

As for the question of humanoid robots when it comes to the elderly care sector, the japanese experience is probably useful to be considered. It also raises the question of social robots which are not always humanoid but have an impact on the way we look at social interactions.

representatives for testing and validating, it is essential to have these users on board at an earlier stage to ensure a meaningful development. While "Diversity and inclusive design teams" are very much needed, getting on board users and considering their experience as an expertise per se is critical and can't be replaced – it is complementary but different approach.

As for the non-technical methods, we would like to underline how much regulation is key and important to ensure a coherent approach. Considering how sensitive a number of issues are in relation to AI, we can't rely only on self-regulation and voluntary guidelines.
When considering education and awareness, it is really important to make a proper mapping and to address different target groups to avoid any gaps. This might require different types of tools outside the spectrum usually used.

should be equally on board.
- It is an area where the human relationship is at stake and where the human interactions shouldn't be lowered down because of AI.
- There might be situation which are potentially highly sensitive and must be handled with carefully, notably when it comes to patients with dementia or cognitive impairments – the informed consent should remain at a core principle.

(2) Autonomous Driving/Moving: there are at least two critical issues linked to the accountability and to the human control, i.e. it is necessary for the users to be able to take control over the "machine".

(3) Insurance premiums: here again the quality of data is critical but the privacy/freedom issue as well. This is an area where a constant dialogue would be needed to avoid any discriminatory approaches (for example on an age basis) and have a comprehensive interpretation of data. There would probably be difficult questions to be carefully considered around the individual responsibility, notably when it comes to health insurance and healthy lifestyles.

There is another area which could also be considered, i.e. smart homes and how much the interrelationship between different kind of systems bases on AI can be at stake in the domestic environment considering it can bring to some sort of abuse.

Typhaine	Beaupérin	FERMA	The Role of AI Ethics (page 2)The definition of an ethical framework is a public policy matter as it relies on a set of values shared by a society. Respect for these values and their practical implementation is of utmost importance to gain the trust of AI users. The use of ombudsmen as public advocates should be promoted to assess the implementation of local norms and whether user information is sufficient and in an intelligible format.Purpose and Target Audience of the Guidelines (page 2)FERMA supports the proposal to set up a mechanism enabling all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis. This is a flexible approach and an opportunity for organisations to demonstrate their commitment to ethical use of AI.However, such Guidelines should benefit EU citizens and EU business competitiveness as well. Therefore, FERMA suggests that the document should explicitly state that the stakeholders invited to voluntarily endorse the Guidelines should include not only organisations established in the European Union, but all organisations that serve EU citizens, businesses and governments, wherever in the world they are based.Most European citizens' personal data is controlled by non-EU businesses, and data is a fundamental pillar for the development and improvement of AI. Having a framework imposing safeguards on AI should apply voluntarily to all businesses operating AI with data from the European Union.	Opt out (page 7, 10)3. Fundamental Rights of Human Beings4. Ethical Principles in the Context of AI and Correlating ValuesThe document refers to an ability to "opt out" of AI decisions. FERMA is concerned about the possible impact on the insurability of a business that decides to opt out from AI decisions (due, for instance, to widespread employee refusal, concerns over data…) when such AI is used by insurers in their underwriting process.Redress issue: "must" (page 10)4. Ethical Principles in the Context of AI and Correlating ValuesThe draft guidelines state that if "harm" is caused as a result of AI, users "must" be redressed. Proposed methods of redress include but are not limited to, monetary compensation and reconciliation. FERMA is raising attention about the impact that such proposals could have upon businesses and their ability to obtain insurance cover when they operate AI tools. Asymmetries of power or information (page 12, 13 and 18)5. Critical concerns raised by AIThere are concerns over situations with asymmetries of power or information, such as between employers and employees, or businesses and consumers. FERMA would like to also raise the issue of asymmetry between businesses and insurers or business-to-business in general. If AI is deployed for underwriting purposes, insurers will have considerably more knowledge than the insurance buyer about how AI has been integrated into the underwriting process and, about how certain conditions will impact the premium as calculated by AI. In this event, the asymmetry of information could potentially benefit insurers in a disproportionate manner as the underwriting process would be increasingly opaque to the corporate buyer.	1 Accountability (page 24)For FERMA, this section is important as it aims to reconcile fundamental human rights and corporate interests. To foster a trustworthy AI among citizens and businesses, FERMA invites the High-level Expert Group to enlarge the scope of risks arising from the use of AI. It should include environmental risks, notably linked to the excessive energy consumption of computing power, but also societal risks, regarding the use of AI by state actors for large segments of the population or through internal changes to business organisations (like the replacement of employee skills, loss of autonomy and management changes when AI and humans must work side by side) as the main business impacts. For all these matters, Boards are accountable and should be supported by all relevant stakeholders in the organisation.FERMA supports the view that risk managers are well placed to analyse the risks related to the non-ethical use of AI, relying on holistic risk management methodology like Enterprise Risk Management.2 Data governance (page 25)FERMA considers that in this section, it will be necessary to identify the rules, frameworks and standards applicable as ethical references for using AI in a business. The main challenge of data governance will be the ability of businesses to measure, correct and explain the differences and deviations between the unrepresentative data which has arisen from factors such as sampling bias and modified, representative data. 3 Design for all (page 25)The analysis of the original data and trends is only the first step of the analysis. The second step of the analysis for a business will be to take a corporate decision that respects the ethical framework while being economically viable.4 Governing AI autonomy (page 25)FERMA supports the view that it is the responsibility of a business to provide users with all the tools to understand how AI is embedded in the services and products offered. Citizens, employees and other AI users should be educated to remain autonomous and independent of AI in their decision-making processes.Risk Management should integrate an assessment of AI in its annual review which would encompass the points listed below:- The possibility to come back to a human-only interaction mode o Was this used when necessary?- An analysis of the deviations in the results of AI.- An analysis and assessment of the impact of having to operate without AI tools, in the event a major issue arose.- Verification of a sufficient level of training for users.5 Non-discrimination (page 25)Informed decision-making processes and the presence of a solid risk governance framework can only benefit the development and use of a trustworthy AI within businesses. A well-identified internal liaison person in the organisation, able to deal with these topics, should put in place a feedback system for issues met by users about biases in services and products.Insurance premiums (page 28)FERMA welcomes the fact that insurance premium is one of the 4 use cases to be developed to operationalise the assessment list.Many Risk Managers are also corporate insurance clients playing a crucial role within their organisations with respect to treatment of complex risks and insurance issues.AI is about opportunities and challenges, one of them being the possible loss of ethical	FERMA is the European federation of 22 national risk management associations. We represent risk and insurance managers active in a wide range of industries (energy, transports, manufacturing, telecoms, financial services…). Our response has been built from the perspective of corporate users of Artificial Intelligence (AI) technology and not only creator/ developers.FERMA welcomes the overall approach taken to establish the first European guidelines on AI ethics. We appreciate the fact that the proposed guidelines are voluntary and built on a set of existing fundamental values, rights and principles. The ethical consequences of inappropriate use of AI are addressed in an objective process based upon recognised and long-standing ethical values. We see the draft as a starting point to efficiently manage the ethical challenges of AI.In FERMA's view, AI should be clearly defined as a technology using a series of diverse techniques (statistics, algorithms, data processing…), upon which rules are coded and programmed to learn without human intervention. The definition should avoid anthropomorphic terms such as "perceiving" and "behaviour", and instead focus on the actual tasks carried out by AI. Such an approach would ensure that AI capabilities would be neither under- nor over- estimated. FERMA remains particularly vigilant and impacted by AI ethics. Vigilant as we see the development of ethical rules as the opportunity to ensure there is accountability in the sphere of AI. Impacted, because we expect the professional practice of risk management to play a fundamental role in the implementation of AI.Indeed, risk managers will have to take the lead on AI topics and analyse all the risks arising from the use of AI within organisations according to different angles, including from an ethical perspective. FERMA supports the view that risk managers are best placed in the organisation to analyse the risks related to the use of AI, relying on holistic risk management methodology like Enterprise Risk Management, which involves conducting a diligent assessment of all possible risks facing the organisation in question. It combines both likelihood and potential impact levels as well as financial exposure on a national and international scale.We currently believe that the main source of risks from non-ethical use of AI are dataset quality, bias and the human factor (error, malicious actions). As for consequences, they are mostly societal (employability of people, discrimination, privacy), environmental (excessive energy consumption) and reputational.Finally, FERMA also draws attention to the implications of AI ethics in the insurance underwriting process and the opportunities and threats of AI technologies for the insurability of organisations.Executive Summary (page i)FERMA argues that the statement: "Given that, on the whole, AI's benefits outweigh its risks" is a strong assertion that deserves at least a transparent and documented explanation justifying it.Glossary (page iv)AI The proposed definition of AI uses several anthropomorphic terms, giving the impression of certain forms of feelings or emotions. The terms "perceiving, reasoning and behaviour" are terms mostly used for living beings.FERMA suggests the following changes to the definition of AI to better

					control over the insurance underwriting process, especially if it is left to the entirely in the control of AI. An ethical debate is necessary to draw a clear line between the opportunities of AI technologies and the threats posed by the same technologies on the insurability of organisations.		reflect its true nature:• AI is composed of algorithms aimed at imitating different cognitive functions like perception, memory, reasoning and learning to reproduce certain competences like organisation, description and information processing. • These processes are performed in an autonomous fashion and involve the processing of complex and unstructured data like images or voices on an unprecedented scale.• Embedded within other technological vehicles, AI can also drive, move objects and perform a series of tasks of various complexity.The discourse surrounding AI technology is impactful upon the public's perception, and thus, we believe it must be accurate and not amplify fears regarding AI. BiasThe definition of bias refers to concepts like general interest and common goods, which are extremely difficult to define, let alone quantify.FERMA shares the view of the High-level Expert Group that the impact on various vulnerable demographics should be assessed in the early stages of the design process through testing and validation.In addition, when datasets are modified to overcome unrepresentative data and bias decisions, FERMA recommends that the original data should be held as a reference to allow the business to assess if their ethical objectives have been met and to constantly monitor the impact of their modifications to the dataset. Moreover, FERMA raises the attention of the High-level Expert Group to the importance of the internal decision-making process in an organisation. We believe it is extremely important for businesses to ensure data subject to AI processing is accurate, of good quality and free from sampling bias. For each subject, the definition of the ethical framework and its granularity needs to be adapted according to the subjects (business to business markets, public to citizens services, governments, trade associations, civil societies…).
Kumiko	Uegaki	Japan Business Council in Europe (JBCE)	• Japan Business Council in Europe (JBCE) appreciates the hard work the HLEG has done to deliver the draft guidelines. We understand the HLEG will continue to work and use the guidelines as a basis for future policy and investment recommendations of the European Commission. • We welcome that the guidelines take a rights-based approach to AI ethics and that the whole text of Chapter 1 is based on the fundamental rights commitment of the EU Treaties and Charter of Fundamental Rights and that these are used as the stepping stone to identify abstract ethical principles. While 'ethics' can sometimes be subjective and their interpretation can differ from one country to another, the "human rights" aspect is clearly defined in the Universal Declaration of Human Rights which is considered as the universal norm. This aspect is important for business because AI systems aren't only made in Europe but are distributed within our international network. We propose that the Guidelines should also point to internationally recognised standards such as the UN Guiding principles on Business and Human Rights (UNGP). In this regard, JBCE recommends to use the guidelines as a basis for discussion at international level and vis-à-vis countries	• Section 4 Ethical Principles in the Context of AI and Correlating Values JBCE supports the principles and related values which must ensure that AI is developed and used in a human-centric fashion. Our member companies do ask, however, for more clarity about "The Principle of Explicability". As opposed to the other 4 principles, 'explicability' is not directly supported by a fundamental right. The description of transparency in Chapter 2 and in the Assessment list is at this moment not sufficiently clear and detailed to provide a good understanding of the principle. We encourage the HLEG to try and better define the principle and/or embed 'explicability' into the other 4 principles. Additionally, a case by case implementation approach would help to understand how each AI system is in line with the "Transparency" requirement. • Section 5.5 Potential long-term concerns JBCE would like to propose to use a different approach to "Critical concerns raised by AI". There is a need to differentiate between, for example, Lethal Autonomous Weapon Systems (LAWS) and other AI applications that relate to privacy, identification and consent. Moreover, within the area of identification there will be different types and levels, as well as different potential	• Section 1-4 Governance of AI Autonomy The Governance of AI Autonomy should be better defined. Care needs to be taken in operational environment to ensure that the human does in fact deviate when necessary, for example, in overriding advice from mostly autonomous driver. Again when talking about 5.4 Lethal Autonomous Weapon Systems (LAWS) In Chapter1, JBCE strongly advices HLEG to address this area at international level. • Section 1-3 Design for all The phrase "Systems should be designed in a way that allows all citizens to use the products or services, regardless of their age, disability status or social status" should be changed to "Systems should be designed in a way that considers usability and accessibility so that the products or services should be inclusive and can be accepted by as many citizens as possible, regardless of their age, disability status or social status." Some systems are designed for specific users and specific applications. For instance, a particular system could be developed for a specific manufacturing process and for employees that have a particular set of skills or expertise. • Section 1-10 Transparency		• "Trustworthy AI made in Europe" JBCE strongly encourages the HLEG to replace the concept of "Trustworthy AI made in Europe" with "Trustworthy AI made for Europe". AI technologies and systems are developed globally, i.e. through companies' research and design centres located in variety of regions and countries, including in Europe, cooperating with each other and their local partners, such as other companies, universities etc.. This applies not only to companies with a Japanese parentage and global headquarters in Japan, but also companies with parentages and headquartered in Europe, as well as in other regions. The AI context in this respect is no different from, for example, the Cybersecurity context or the IoT context or the 5G context or the blockchain context. Likewise, international cooperation needs to take place also at the level of ethics, standards of interoperability, international regulatory cooperation and common investment on R&I. As for R&I, JBCE member companies have been participating in the EU's relevant programmes, including Horizon2020. This cooperation should be reinforced in new Horizon Europe programme in AI area. • Glossary "Bias"

			with a similar approach to AI ethics and share the EU's main goals.	users of AI. JBCE strongly supports a case by case approach both concerning AI applications and different level of AI used within one field of application, as well as in relation to the potential risks that are associated with them. When talking about 5.4 Lethal Autonomous Weapon Systems (LAWS) In Chapter1, JBCE strongly advices the HLEG to address this area at international level.	The term "development processes" which is used in the fifth line of the page should be clarified. If the term refers to the design process for system software, then it could be difficult for companies to be transparent about such design processes. The process that a company follows while designing its system software is one of the important factors that differentiate the company in question from its competitors, and thus disclosing these types of processes could be difficult.		The definition "Bias" should be more emphasized on which type of data and how this data is collected so that providers of AI systems could be liable. --- The JBCE is a European association representing over 80 multinational companies of Japanese parentage in the EU policy discussions. Our members are active in Europe across many sectors, including digital, information and communication technologies, electronics, automotive, pharmaceuticals and chemicals. JBCE acts as a bridge between the EU and Japan to strengthen ties and demonstrate to European decision-makers the contribution of Japanese companies in Europe.
Francesca	Fanucci	European Center for Not-For-Profit Law (ECNL)	We welcome the provision of the AI draft Ethics Guidelines establishing that, "A mechanism will be put in place that enables all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis" and that such mechanism will be "set out in the final version of the document." (p. 2). We recommend that such mechanism should include a periodical review of the number of stakeholders that have signed up to the Guidelines and should require signatories to report on a regular basis: 1) how they have concretely implemented the Guidelines; 2) how they are assessing the effectiveness of their own initiatives to implement the Guidelines.We also recommend including the need to hold a periodical review of the Guidelines themselves in order to assess the challenges encountered in its implementation and the potential need to update their content. This review should be conducted in the form of an extensive public consultation engaging all relevant stakeholders, including civil society organisations.	In their approach to AI ethics, the Guidelines explicitly refer to "the fundamental rights commitment of the EU Treaties and Charter of Fundamental Rights as the stepping stone to identify abstract ethical principles, and to specify how concrete ethical values can be operationalised in the context of AI" (p. 5). We recommend that a specific reference should also be included to the European Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR) as an integrating part of the interpretations of the EU Treaties and the Charter on Fundamental Rights (Charter). Art. 6 (3) of the Treaty of the EU explicitly acknowledges that the fundamental rights guaranteed by the ECHR "shall constitute general principles of the Union's law." Furthermore, Art. 52 (3) of the Charter stipulates that the meaning and scope of the fundamental rights that correspond to those guaranteed by the ECHR of the Charter "shall be the same as those laid down by the said Convention." The meaning and the scope of the guaranteed rights are further elaborated by the case-law of the European Court of Human Rights and of the Court of Justice of the European Union, which can also help assess and resolve existing concerns in the context of AI. I. 3. Fundamental Rights of Human BeingsWe welcome the explicit inclusion in the Guidelines of "freedom of assembly and association" among the "rights are particularly apt to cover the AI field" (p. 7). With regard to these two rights, we recommend adding a specific reference to the impact that AI systems may have on civic space, since the development and implementation of AI tools can either promote or challenge the enabling environment for civil society and its organisations. I. 4. Ethical Principles in the Context of AI and Correlating Values The Guidelines acknowledge "the potential of unknown and unintended consequences of AI" and therefore advise "the presence of an internal and external (ethical) expert […] to accompany the design, development and deployment of AI" (p. 8). We suggest advising to also include the presence of internal and external human rights experts and civil society organisation representatives, who can identify potential impacts on specific fundamental rights and freedoms protected by the international standards. I. 5. Critical concerns raised by AI For all concerns described in paragraphs 5.1 - 5.5 (p. 11-13), we recommend referring to	Chapter II.2: Technical and Non-Technical Methods to achieve Trustworthy AI The Guidelines emphasise the idea that, "compliance with law as well as with ethical values can be implemented, at least to a certain extent, into the design of the AI system itself" and that, "this also entails a responsibility for companies to identify from the very beginning the ethical impact that an AI system can have, and the ethical and legal rules that the system should comply with" (p. 19). We recommend adding the words "and fundamental rights" after "as well as with ethical values" in the first sentence and amending "the ethical impact that an AI system can have" with "the ethical and fundamental rights impact that an AI system can have" "in the second sentence. Furthermore, where the Guidelines recommend that, "Organisations should set up an internal or external governance framework to ensure accountability. This can, for instance, include the appointment of a person in charge of ethics issues as they relate to AI, an internal ethics panel or board, and/or an external ethics panel or board. Amongst the possible roles of such a person, panel or board, is to provide oversight on issues that may arise and provide advice throughout the process." (p. 22), we recommend including the appointment in the panel or board of external/internal experts on human rights issues and representatives of civil society organisations, with the same roles and responsibilities.Most importantly, this part of the Guidelines should explicitly include the need for developers and governments to conduct publicly accessible and expert-informed human rights impact assessments in all stages of the process, during development, at regular milestones, and throughout the use of each AI-based systems and services to the public. The specific purpose of these assessments is to identify risks of rights and freedoms-adverse outcomes – not just potential infringements of ethical values - and develop appropriate measures to avoid and mitigate those risks.Human rights impacts assessments should be conducted as openly as possible and encourage active engagement of beneficiaries as well, including civil society organisations. Civil society organisations should also play an active role in monitoring the use of AI systems after implementation, to ensure that each AI system is effectively being used as originally intended.The main	Based on the above-mentioned considerations, we recommend reformulating and adding the following questions in the assessment list as follows:1. Accountability:• Has an Ethical and Human Rights AI review board been established? A mechanism to discuss grey areas? An internal or external panel of experts? Are human rights experts and civil society organisations invited to be part of it? (p. 25)• Is a publicly accessible and expert-informed human rights impact assessment and evaluation in every step of the AI process in place?• Are all persons involved in the development and assessment of the AI systems adequately trained with respect to applicable human rights and freedoms norms and made aware of their specific responsibilities?5. Non-discrimination:• What are the sources of decision variability that occur in same execution conditions? Does such variability affect fundamental rights or ethical principles? How is it measured? Are human rights experts and civil society organisations involved in the measurement of such variability? (p.25)	Finally, the Guidelines should make a specific and stronger call for extensive public consultation and dialogue, not just with the relevant stakeholders but the public at large and civil society organisations. The EU Commission is committed to listening more closely to citizens and stakeholders, broader public, as part of the Commission's Better Regulation Agenda. This includes opening up EU and national policy and law-making and listening more to the people it affects. Quality of dialogue relies on evidence and a transparent process, which involves the public and stakeholders (for example, businesses, public administrations, civil society and researchers) throughout the process. This can help developing and implementation of the AI systems become more transparent, accountable, inclusive and effective, with aim to share and discuss information and promote the responsible use of the AI. In addition, the Guidelines should encourage its signatories to promote digital and information literacy programmes for the public and civil society to enable them to understand AI, enjoy its benefits and minimise the risks arising from it.

				the need to apply the "three-part test" developed by the European Court of Human Rights to strike an appropriate balance between fundamental rights and their potential limitations/infringements caused by the development and use of AI systems. The three-part test requires that a limitation to right/freedom should be prescribed by law, serve a legitimate purpose and be necessary/proportionate in a democratic society. This is particularly important given the EU and its Member States are bound by the Court's interpretations and case law when implementing human rights and freedoms.	findings of any risk assessment process, identified techniques for risk mitigation, and relevant monitoring and review processes should be made publicly available.In addition, all persons involved in the development and assessment of the AI systems should be adequately trained by human rights experts and relevant civil society organisations with respect to applicable norms on human rights and freedoms and should be made aware of their specific responsibilities.		
Arthur	van der Wees	Arthur's Legal	Although the draft on Ethics Guidelines for Trustworthy AI has set the scene by providing insightful information on the purpose of the Guidelines, Arthur's Legal is of the opinion that it may be of added advantage if the final draft of the guidelines reflects on how society has evolved and adapted to the introduction of new technologies. Electricity, the steam engine, automobiles are some of the early examples of General-Purpose Technologies (GPTs) i.e. technologies that act as a major driver in transforming an economy. As a result of modernization and innovation, GPTs like the computer and the internet were invented which ultimately led to the Information Revolution. However, the technology which is currently taking society by storm is Artificial Intelligence (AI). Within a short period of time, AI has become an indispensable part of our daily lives, as we use it many forms: as apps on our smartphones, programmes on our computers, smart cars and smart home appliances. AI has the potential to make our daily lives more comfortable (remotely operated domestic appliances), sustainable (smart meters) and safer (autonomous cars, cctv). Moreover, AI has transformed industries like finance, banking, healthcare and the like. It is said that the adoption of AI could increase productivity by 40% in all major industries by 2035 as it will ensure that people make use of their time in an optimal manner. Just like every coin has two sides, so does AI. While AI has the capability to make life easier, it also has the potential to make life a lot more dangerous and complex. That said, it is pertinent to note that when cars were first manufactured on a large scale in the 1900s, the invention was hailed as a wave of the future but was also viewed with circumspection. Several concerns were raised regarding passenger safety, road traffic management, air pollution that might result from emission from the car etc. However, rather than shunning the use of cars altogether, pragmatic solutions were devised. Legislations mandated the use of seatbelts, traffic lights regulated traffic and vehicle certification systems verified whether emission standards were complied with. If history has taught us one thing, it is that regulation is better than prohibition. Therefore, while dealing with AI, it is important that we gauge the challenges that	Before diving into the requirements for trustworthy AI, it may be beneficial to provide use cases describing situations that may arise in the future in combination with certain questions in order to set the tone of what the user is likely to expect in the document. Not only will this increase the readability of the guidelines, but it will also make it more relatable. Some use cases based on the themes highlighted by the High-Level Expert Group are as follows: 1. HEALTHCARE DIAGNOSE AND TREATMENT According to the World Health Organization, 300 million people around the world suffer from depression. The type of depression varies from person to person with some experiencing short-lived emotional responses while others face more serious health issues. Alarmed by the spiraling rise in patients suffering from depression, Woodland Medical Centre decided to make a change. The medical center teamed up with an AI startup that provided behavioral healthcare solutions by using smartphones to diagnose mental health of patients. On the basis of a wide variety of factors including phone activities, typing speed, physical movement, social media usage and interests, the application could create an image of the mental health of a patient. Additionally, a separate AI backed chat-bot was also provided for as Woodland Medical Centre believed that talking about their problems could help alleviate the anxiety of users to a certain extent. The app later collates the data from different sources on the smartphone and categorizes the users based on the level of depression that the user may be at. Moreover, the data of the users who seem critical would be shared with doctors and psychiatrists to allow them to make a separate assessment based on the data available. In such a scenario, the following questions that need to be looked into are: • Is the Medical Centre transparent about how it collects and shares data? • Are there sufficient safeguards to ensure that the application is not hacked? • Are there measures to see how the app reached its decision? • Who will be responsible if something goes wrong? • Is the app continuously upgraded to ensure reliability? • Have provisions been made to eliminate bias?	The second chapter of the Guidelines provides a detailed list of technical and non-technical methods that need to be factored in to create trustworthy AI. As already mentioned in the Guidelines, the list is not exhaustive and hence, Arthur's Legal would like to put forth some additional suggestions: • Continuous Updates From the inception of an AI device, it should be continuously updated in order to to reduce vulnerabilities. The Guidelines currently talk about how AI device should be created but it does not deal with the life of an AI device and how it should maintain checks and balances to prevent its vulnerabilities from being exploited. While training a system and after its deployment, it should be constantly monitored so as to detect any flaws or vulnerabilities that may arise during its life cycle. This process should be followed up with suitable updates to make the system more secure. • Data Minimization Given that AI devices thrive on the data that they collect or that is provided from an outside source, AI developers, manufacturers and service providers should take all the necessary measures to evaluate the quality, nature and amount of personal data that it collects and uses during the training and development stage. Enforcing this principle will require developers to objectively consider the intended areas of application of the AI device and facilitate collection and usage of data for the intended purpose. While it may be difficult for the AI developers, manufacturers and service providers to establish in advance the kind of information that they may require for the development of an algorithm, the principle of data minimisation will require them to make a continuous assessment of their actual requirements while also weighing them against the right to privacy of the users. Synthetic data i.e. data that is generated by a computer and not human by mimicking real data can also be used to train AI models. • Data Encryption Encryption can and should be used as an effective tool for ensuring the security of data. It has been used in the past for various technologies, it should remain a part of the design of AI technologies, applications so that in the event of a hack, information is	In the final chapter of the Guidelines, the High-Level Group has provided an assessment list based on the 10 requirements for trustworthy AI. Through this document, we would like to provide the following questions that could be considered while drafting the final guidelines. i. How will liability be assessed/determined if something goes wrong? Considering the complexities of AI, unforeseen consequences are inevitable. While monetary compensation and reconciliation have been provided for in the Guidelines, there may be incidents with serious implications which involve human lives. In such a scenario, how will liability be determined? ii. Who has the burden of proof? Given the multi-party involvement in creating AI, who will be the burden of proof rest on? For example- In the case of autonomous vehicles, will it be on the sensor hardware manufacturer be or the developer who created the code? Would the car manufacturing company be entirely responsible? iii. Need for a start button? The guidelines currently question the provision of a "stop button". The need for a kill switch has been discussed in the past with respect to AI in order to prevent a unforeseen negative incident or to shut down the AI in case a major error is detected. In such a case, will the system automatically start to function after the issue is corrected? Should humans be given the right to determine when the said system should start? Within this notion of interference (start, stop and kill buttons) the consequence of that needs to be considered as well, such as reversing analytics and decisions, and perhaps even deleting certain parts of the evolved/resulting AI (or at least containing it to mitigate further damages and the like). iv. As already mentioned, the Guidelines provide accountability mechanisms ranging from monetary compensation to reconciliation. Is there an enforcement authority that will determine which recourse should be taken? If not, then who will decide? v. Who owns the algorithms and benefits from them? Most AI glean the insights and give meaning from millions of datapoints. i.e. from millions	Arthur's Legal is of the opinion that the Guidelines can be made more user friendly and easy to approach given that artificial intelligence is becoming more mainstream. We would be happy to help out with that as well. Not only should the guidelines focus on the obligations of AI developers, manufacturers and service providers but it should also highlight the rights of AI users and how they can enforce the said rights. By doing so, AI developers, manufacturers and service providers will be aware of what is expected of them when creating AI while users will be more aware of what they are entitled to expect when they use such technologies, thereby ensuring that the Guidelines are implemented in a more holistic and seamless manner. Thank you for this opportunity to contribute. We are looking forward contributing more, either on paper, face to face and by other communication means.

might be faced in the future with respect to AI and formulate a holistic and human-centric approach that will not only make AI easier to use but also safer.

While the High-Level Expert Group on Artificial Intelligence has provided a comprehensive and coherent draft on Ethics Guidelines for Trustworthy AI to address various concerns, Arthurs Legal would like to put forward certain recommendations that may be looked into in order to ensure that an all-inclusive, future proof approach is taken while formulating the final draft of the guidelines.

2. AUTONOMOUS DRIVING/MOVING

The field of autonomous vehicles has turned science fiction into reality. In order to enter the market of such vehicles, the German company AXYZ started manufacturing its first series of driverless luxury cars. In keeping with its range, the car does not make any provision for manual controls like a brake pedal, steering wheel or accelerator. The cars have cameras, a facial recognition system, short-range radars and long-range radars along with other sophisticated assistance systems that allow the car to reach its destination. Several ethical questions come to the forefront in such a situation, including:

• Is there stop button that may allow the rider to intervene?
• Will the systems in the driverless cars be in a position to make a decision between sacrificing itself or saving 30 school children that may be standing on the road?
• Would the answer above be different if it involved a jaywalker?
• Are the systems safe from hacking and unauthorized access?
• In case of an accident, who will be liable?

3. PROFILING AND LAW ENFORCEMENT

After a survey by the United Nations revealed that the crime rate in country ABC had doubled in a period of five years, the crime branch of the country decided to tie up with a data analytics firm 247Analytics. As a part of the engagement, the crime branch was required to share various forms of data of its residents including addresses, phone numbers, court filings, previous criminal records (if any), criminal database, social media data etc. It was expected that by sharing such data, the crime branch could engage in predictive policing i.e. use an algorithm that could predict, based on the data imported, whether certain individuals would engage in criminal activities. While the intention of the crime branch is in the right place, the following questions need to be considered:

• Given that there will be large amounts of data which will be highly confidential, what security measures will be put in place to prevent hacking?
• Will the residents be informed about the kind of data being collected and for what purpose?
• Is there a strategy in place to avoid biased decisions based on racial backgrounds or level of income?
• Has a fall back plan been formulated?
• Is an oversight mechanism in place to ensure that fundamental rights of residents are respected when weighed against the crime branch's objective to reduce the crime rate?

4. INSURANCE

ABC Insurance company, in its objective to deliver affordable healthcare to a greater population in a structured manner created an algorithm that determine the insurance premium in a fast and efficient manner based on certain parameters. For every individual, the algorithm would assess the nature of their job, possibility of injury or not easily available to unauthorized parties. With respect to AI, it may be advisable for AI developers, manufacturers and service providers to use homomorphic encryption as it will allow them to use data to perform operations without having to decrypt it. As a result, systems may be allowed to use sufficient data for training with lowered risk of data breaches. Apart from homomorphic encryption, stakeholders must be encouraged to take active measures to explore other methods for encrypting data in an effective and efficient manner.

• Collaboration with Experts and Universities
It may beneficial for AI developers, manufacturers and service providers to set up committees of experts from different fields to have a diversity in perspectives for developing AI in an ethical manner. Academic institutions may also be consulted given their social and public interest in designing human rights based and ethically oriented AI applications. Moreover, these universities may be able to initiate a dialogue with different stakeholders in the AI ecosystem that may have otherwise been reluctant to participate in such a dialogue.

• Enforceability
While ethics play an integral role in creating the parameters for conduct, while dealing with complex technologies like AI, it is imperative that this conduct translates into actual behavior while also ensuring Accountability, Responsibility, Liability. The Guidelines also need to focus on Recourse and Remedies, including without limitation how to protect society & economy, individuals & organizations, public sector & private sector, professionals & SMEs, against the risks, mistakes, malicious acts, intentional or nor, and the impact, effects and consequences thereof. While dealing with AI, it is important that the entire AI ecosystem is taken into account. The Guidelines are structured and architected around AI developers, manufactures and their obligations, however , it should also throw light on how users can enforce their rights where the said developers and manufacturers fail to hold their end of the bargain as a result of which the users suffer.

• Burden of Proof
AI consists of a highly complex value chains, linking different hardware and software components together, communicating with one or more networks and other devices. In such a situation it will extremely onerous for the consumer to hold one party liable in such a non-linear and multi-dimensional web. The burden of proof must be on the one that makes, maintains, controls, uses and deploys AI – which may not be one party. It will be unreasonable to expect society, a person or organization downstream to prove that AI was (part of) the cause.

These difficulties are worsened because, whilst trying to comply with this burden of proof, persons, organisations and society in general are often to a great extent dependent on the information provided by the AI vendor or developer. This information is often (deliberately) kept very limited and vague.

Furthermore, a more contextual approach has to be taken while dealing with this

of individuals? That said, how should the wealth that is generated by such AI be distributed? Additionally, with such diverse and non-linear participants, who will be the ultimate owner of the algorithm?

vi. The Sustainable Development Goal of inclusion may be overlooked or further dispossess the marginalized by capturing the value that inherent in the data about them or even produced by them (as citizen-generated data). How can this be weighed against the interest of the AI developer, manufacturer and service provider?

vii. How might the algorithms that power AI be made more democratically or inclusively available?

viii. Like open source software, if the source code governing the AI are made open source, AI may be used for diversified purposes by the general public. Is there such as think as Open AI? Should there be?

ix. Should a certification system be established for AI that indicates their ethical adequacy? Who could be the certifying authority?

				permanent disablement during the job, monthly income, ethnicity, citizenship, permanent address and the like. While affordable healthcare is in the interest of individuals, the following issues need to be assessed: • Were the individuals made aware of how their insurance premium was being determined? • Would there be an alternative if they refused to such automated decision making? • Can the individual challenge the decision made by the AI if they believe that the decision was biased? • If the allegation is proven to be true, is there a mechanism to correct the error in the algorithm to prevent similar decisions from being made in the future?	aspect as it depends on the functionality of the AI, the purpose of it, the question whether it only analyzes, or also makes decisions or even further executes those as well, the domain it is deployed, the detrimental impact it can have, et cetera. To an extent, this is reflected in Article 25 (Data protection by design and by default) and Article 32 (Security of processing) of the General Data Protection Regulation. • Right to Object While the Guidelines state that humans should be made aware of times when they are interacting with AI, however, it is equally important that individuals are informed about their right to object to processing of their data by such systems that have an impact on their behaviors, choices or opinion. When automated decisions are made, efforts should be made to implement measures that protect the fundamental freedoms and privacy of individuals. • Custodianship: Trusted Third Party Deposit There are several reasons and situations (*) why it makes sense to consider setting up an European AI trusted third party ecosystem where AI is deposited in a secure and safe way and kept by such independent custodians, and which – for instance only after (A) approval on a case by case basis` by the respective AI developers, manufacturers or service providers, (B) its conditional pre-approval or (C) a court order or arbitrational decision (such as but not limited to a binding advice) – gives access to the particular AI for independent auditing, due diligence or other forensics. This will increase trust and transparancy for society and economy, and the compliance and accountability of the AI by the respective AI developers, manufacturers and service providers, while appreciating and protection the (possible) competitiveness and business/value models thereof. (*) Such reasons and situations could include (A) being part of critical infrastructure, vital systems or essential services (NIS Directive), processing personal data (GDPR), or (B) being part or otherwise relevant for incidents, accidents, disputes, court cases, arbitration, settlements and the like.		
Andreas	METZGER	NESSI (The European Technology Platform dedicated to Software, Services and Data)	There are several references to "technology", but explicit reference to software is absent. The software implementation of AI algorithms is fundamental to the achievement of the vision for AI. The governance of software technologies and suitable software engineering methodologies for AI are challenges which should be explicitly addressed. Also the new class of self-learning and self-adapting systems – empowered by AI – should be considered.	In applying the law to fast developing technologies such as AI, regulation may lag behind technical development, making it difficult to anticipate how systems might be treated judicially. Recent discussions of legal personhood for robots and AI are clearly a long-term approach, but for the foreseeable future humans will have ultimate responsibility for the decisions taken by AI systems. This implies the need for a framework of software and systems engineering requirements for determinable legal liability. It is also important that developers of AI systems have a means to assess the ethical dimensions of their design choices, given the potential for such systems to disrupt and damage the legitimate interests of citizens. AI systems need to be considered as complex multi-stakeholder systems in which the designers as well as system operators and end-users are moral agents who must share the legal and ethical	Data governance gets plenty of attention; equally, governance of software technologies needs to be addressed. Underpinning this is a need to guarantee the predictability and governability of self-adapting software system and architectures.AI-based self-adapting systems help master the complexity, dynamicity and uncertainty entailed in developing software systems. By learning at run-time, they can handle situations that cannot be anticipated at design time, due to incomplete knowledge and uncertainty about the system environment. However, such online learning can create difficulty when developing, debugging and testing systems that can self-adapt, e.g. determining causality and liability for autonomous actions and decisions. The lack of transparency regarding how a system works, and who or what is responsible for the resulting output, can raise concerns. Explainable AI, including algorithmic	Certification of much AI software technology is a huge problem: when an AI-based system is self-adapting by learning, there is never a stable operational version to certify.One option for governance today is to use human oversight; in future, AI systems may be required to use automated failsafe mechanisms, which will rely on monitoring systems that can detect failure modes, and will execute remedial actions to restore safe operations even in situations that have not been previously defined and analysed.AI-enabled software systems introduce a host of other quality assurance and security challenges (e.g. AI/ML components may introduce vulnerabilities that cybercriminals can exploit). This new situation is likely to require a new programming paradigm that will enable the smooth integration of e.g. (by nature non-deterministic) ML components and interfaces in a readable and debug-able manner. AI/ML techniques themselves could	There is a lot of focus on data and AI, because the latest AI algorithms and applications are typically data hungry. However, without software there is no AI. Yet in the draft guidelines, the word "software" appears precisely once, on p17, whilst "data" is everywhere. Algorithms get mentioned several times, but having ethical trustworthy algorithms is not sufficient; it is the software instantiation of an algorithm that actually gets executed. Self-adapting software in particular is a major challenge. The governance of software technologies for AI thus should be explicitly addressed.

				responsibility for the outcomes.AI will lead to a proliferation of self-learning and self-adapting systems, which are non-deterministic and thus can behave in ways the developers never imagined. How can designing in controls to ensure ethical behaviour be respected by a self-adapting system? Such systems can also be used in ways the developer never imagined. The guidelines say that "the principle of justice also commands those developing or implementing AI to be held to high standards of accountability". Where does the accountability of the developer end, and that of the user begin? Many tools can be used for good or harm – is that the responsibility of the designer or manufacturer of the tool, or that of the user of the tool?	transparency, should allow verification, even by non-experts, and can contribute to discovering errors or biases that otherwise would have been left unnoticed.	potentially support software development with new validation and formal verification methods, helping create high-quality, less vulnerable and more secure code.
Natasha	Duarte	Center for Democracy & Technology	The Center for Democracy & Technology supports the High-Level Expert Group (HLEG)'s efforts to develop guidelines for trustworthy AI and appreciates the opportunity to comment on this draft. In particular, we commend the group for affirming a rights-based approach to governing AI, for moving beyond the development of principles, and for acknowledging the need for a context- and domain-specific implementation of the values discussed in these guidelines. While we agree that trustworthiness is a key objective for any system, the HLEG must also acknowledge the limitations of current methods for mitigating bias in machine learning models. In many contexts and applications, truly trustworthy AI remains hypothetical. Moreover, trustworthiness depends not only on the ethical purpose and technical robustness of the model or application but also on the governance of the entire societal context or legal system within which an AI application sits (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3265913). We recommend that the HLEG place greater emphasis on (1) the importance of mechanisms and processes for continually interrogating and challenging AI systems from both the inside and the outside and (2) the importance of assessing the entire system (including underlying policies, laws, and human-technology interactions) that surround the AI. The draft guidelines avoid the pitfall of relying on ethics alone as a solution to mitigate the potential harms of AI systems. Many industry actors have embraced ethical codes, but they rely too heavily on the ability of a particular company, or individual or team within the company, to weigh the ethics of a system and decide what is best for everyone it will affect. As AI Now wrote In its 2018 report (https://ainowinstitute.org/AI_Now_2018_Report.pdf), "Ethical approaches in industry explicitly ask the public simply take corporations at their word when they say they will guide their conduct in ethical ways," and "ethical codes may deflect criticism by acknowledging that problems exist, without ceding any power to regulate or transform the way technology is developed and applied." Ethics may also fail to address how an open-source technology may be used by others who are not bound by a particular ethical code. Instead, the HLEG is right to point out that the	Once again, we commend the HLEG for re-affirming the EU's "rights' based approach to AI ethics." However, when fundamental human rights are translated to ethical "principles and values" to govern AI, it is likely that different stakeholders and decision makers will apply the principles differently. This is particularly true of beneficence ("do good"). We have found that between civil society and industry, and even among industry actors, beliefs about what technologies or designs benefit society can diverge widely. Moreover, in many applications, it may be far from clear how the values articulated by the HLEG should be balanced against each other. For example, one focus on AI fairness/ethics research in the EU and the US is on how to create recommender systems (for news, entertainment, trending topics, jobs, etc.) that are more equitable with respect to the diversity of publishers that are able to reach an audience or that promote a more diverse (or less polarized) information diet (https://piret.gitlab.io/fatrec2018/program/). This work advances values such as non-discrimination and beneficence (they arguably "do good" by reducing polarization or disinformation in news dissemination). However, it could also be seen as interfering with human autonomy by nudging people toward content that they wouldn't otherwise choose or suppressing the effects of majority preferences. It is likely that different stakeholders would balance these values very differently. We recommend that HLEG include more discussion of the right to an effective remedy (or redress) in this section. Specifically, "explicability" should serve not only to inform citizens about the existence and operation of AI systems to build trust but also to facilitate effective appeal and remedies. This is particularly important given that, no matter how much due diligence is performed, AI systems will continue to make mistakes.	Governance: The guidelines discuss the need for governance of both data and "AI autonomy," but they are missing a discussion of governance that extends to the larger system or context within which the AI is deployed or with which it interacts. For example, automated decision systems are being developed and deployed to replace or (more often) assist with human decision making. When researchers and civil society study the ethics or fairness of these systems, we often find that the problems are not necessarily (or not only) within the AI or even the deployment and governance of the AI but deeper within the pre-existing system or structure. No matter how technically robust or ethically designed an AI system is, it will not be trustworthy if it is designed to execute decisions in a system that itself is untrustworthy. Privacy: As the HLEG acknowledges, it is critical to test AI systems' performance on different subgroups, particularly vulnerable and minority groups, in order to identify and mitigate discrimination. This may require the collection or inference and use of sensitive characteristics. Collecting or inferring this information while maintaining appropriate privacy protections raises challenges without easy answers, and this will be a critical area for legal and technical analysis over the next few years. Privacy laws are critical but should not become a barrier to assessing AI for discrimination and protecting vulnerable groups. The Commission should consider providing guidance, with input from affected communities, on ways to collect sensitive-characteristic test data while complying with the GDPR. Documentation and ethical constraints in open AI: Machine learning models and training data sets are often made available for anyone to use and incorporate into their own project or product, or train using their own dataset. In order to ensure values such as safety, non-discrimination, and robustness, it is not enough for the original developer or data collector to hold themselves to those values. They must consider how their designs or datasets may be used and iterated on by others, including malicious actors.	The draft proposal includes many helpful questions for the assessment of Trustworth AI. In addition to those, please consider including the following questions: Design for all: Do people have a non-AI alternative or substitute for the system or service? Considering the availability of alternatives, or lack thereof, may inform the degree to which users of an AI system do so out of need, versus choice. Safety: Have the effects of the risk mitigation or management plan been tested? From a trust perspective, proof of the effectiveness of risk mitigation and management is an important metric. Has there been an assessment of the potential to improve the risk mitigation or management measures? Where such measures have been tested, trust in systems could be enhanced by evidence of learning from past iterations. Have the effects of interactions between multiple AI systems been identified? As AI systems become more ubiquitous, they may be more likely to interact with each other and produce unanticipated effects. Attempts to at least identify the potential risks associated with such interactions could be helpful in an assessment of trustworthiness. Transparency: Purpose- Is/are the system(s) being used as intended? As systems are deployed and used, it will be important to understand how, in what contexts, and for what purposes they are actually used. What measures have been taken to limit unintended uses? Where unintended uses are known, or anticipated, measures designed to limit either the use or the effects of unintended uses could help to ensure that fundamental rights are not inadvertently infringed. What impacts might the system have on the fundamental rights of the intended users? Given the rights-based approach for assessing Trustworthy AI, an impact assessment through the lens of fundamental rights should be included.

			development, deployment, and use of AI must "respect fundamental rights and applicable regulation." Similarly, we support the HLEG's effort to move beyond principles to more concrete guidance. However, as the HLEG acknowledges, general high-level guidance cannot address the context- and domain-specific challenges, ethical considerations, or rights implications of an AI application. The European Commission is setting an important example for the rest of the world by affirming that competitiveness in AI need not and should not come at the expense of human rights or ethics. While we support the goal of trustworthiness, we caution that for many AI models that have been or are currently being developed, the trustworthiness of the model itself remains hypothetical. Research into methods for removing or avoiding harmful biases in machine learning models is progressing (http://proceedings.mlr.press/v81/), but we are still far from solving these problems (https://www.omidyar.com/insights/public-scrutiny-automated-decisions-early-lessons-and-emerging-methods). Thus, rather than "trust in technology," the current moment calls for an emphasis on trustworthiness in the systems within which the technology is used and the processes that govern its use.		Researchers have come up with ideas for ensuring that open models (https://arxiv.org/abs/1810.03993) and open data (https://arxiv.org/abs/1803.09010) are not misused, including models for documenting the robustness and domain of the model/data and methods for putting fairness constraints into the code (https://dl.acm.org/citation.cfm?id=3287588). However, it will take a multipronged approach and continuous auditing to ensure that open models are not used in untrustworthy ways. Traceability and auditability: These are necessary characteristics for trustworthy AI. However, the draft guidelines over-emphasize the function of transparency for facilitating "laypersons'" understanding of "the causality of the algorithmic decision-making process and how it is implemented . . ." and " the laymen's acceptance of the technology." First, the "layperson" distinction may be misplaced, since even someone with deep expertise in machine learning will not inherently "understand" how every model works and will benefit from auditability. Second, even if useful and meaningful explanations of AI systems can be developed, the people who are affected by those systems should not assume the burden of truly understanding how they work, or how a particular automated decision causes a particular effect on a person's life. For the average person, the choice to an engage with an AI system in some way is unlikely to be truly optional. Thus, "acceptance" of the system should not be presumed from the existence of an accessible explanation of the system. Instead, a primary goal of transparency should be to facilitate appeals and redress when an AI system does something wrong. Stakeholder engagement: The draft guidelines correctly identify the need to ensure the participation and inclusion of stakeholders in the design and development of AI systems that will impact them. However, the guidelines do not elaborate on potential ways to do this. Stakeholder participation is a value that is often identified but rarely operationalized. HLEG could add value by developing concrete recommendations for engaging stakeholders, including the adoption of processes, such as hearings and town halls, where community members can weigh in on AI applications being considered for deployment in their communities.	What impacts might the system have outside of the intended group? Since AI systems may be able to impact non-users, an impact assessment for the potential of a system to affect their fundamental rights should also be part of the overall assessment framework. When systems make decisions impacting people other than the user, such as in autonomous driving systems, are the criteria for balancing risks and benefits to the public communicated? The general public should have a means of accessing information regarding how AI systems will measure and weigh the risks imposed through their use.	
Erin	Green	Conference of European Churches	While this section does a generally good job of establishing context for the guidelines, it requires further consideration and development on a few points. • We challenge the confidence that "no legal vacuum currently exists" (page 2) with respect to European AI regulation. We urge extreme caution on this front considering the flexibility of legal interpretation and the rapidly changing AI landscape that will ultimately subvert existing regulation and legislation. For example, how will regulation handle shifting intellectual property rights,	• There is a tension throughout the document between the individual and the common good that requires clarification in the final version. At some moments, the individual human is at the centre of concern, at others individual wellbeing and the common good are given equal footing (e.g., the guidelines say that AI should "improve individual and collective wellbeing"). The guidelines mention the importance of the treaties in resolving the tension between the individual and the common good (page 8), but this needs to be dealt with in a much	Remarks for this section are already well covered in similar comments for the other sections of the draft guidelines.	• The section on accountability requires some further clarification, especially with respect to the kinds of accountability that are at stake in developing Trustworthy AI. For example, the guidelines do not make a clear distinction between moral, legal, financial, and technical accountability for these technologies. There is a further need in this section to address the ever-present tension between accountability and the desire for secrecy for reasons of competitiveness or national security. Trustworthy AI demands, in part, full	The rise of robotics and AI is an important concern for the Conference of European Churches and its constituency. We appreciated that the Commission and the AI HLEG approached the ethical challenges of AI as an ongoing process that requires a diversity of stakeholders. We are grateful that the European Commission has taken up this work within the scope of the Article 17 dialogue, and hope that it will continue. Preparation for this consultation and the related dialogue seminar, however, did not provide ideal timing or opportunity for

civil rights, and complex liability issues as new technologies emerge that do not fit well into the scope of current regulation?
• It is impossible to speak of "the goal" of AI ethics in the singular—the unresolved debates in the AI HLEG speak directly to this. We also challenge the notion that AI is a "scientific discipline", when it is so clearly an interdisciplinary pursuit drawing on an extensive range of research traditions including linguistics, developmental psychology, anthropology, and social sciences, among others.
• We would argue, instead, for an approach to ethics that decentralizes the individual human and seriously considers the ethics of community life, society, the common good, as well as ecological concerns in light of global catastrophic climate change. Such an approach does not find a home within these guidelines, but we are hopeful that "the beginning of a new and open-ended process of discussion" will inevitably take up these essential perspectives.
• We argue for argue for a more expansive understanding of stakeholders to include all those passively or actively impacted by—not just directly developing, deploying, or using—AI. Passive and hidden applications increasingly shape life in Europe and beyond, despite the call for transparency later in the text. These include applications like vehicle-to-vehicle communication, traffic management, surveillance and facial recognition, which touches the lives of many who fall out of the prescribed stakeholders group. We urge special consideration for minors, who have decisions made on their behalf about their interaction with these technologies. In this sense all Europeans and all who cross its borders—virtual and real—are stakeholders in this process.

more robust way. The Conference of European Churches recommends re-examining the relationship of the individual human to its context and clarifying how these correspond in the final guidelines.
• The section on vulnerable demographics should include women, refugees, Indigenous and traditional peoples. Also, recent examples of bias in AI show that racialized and queer people are especially vulnerable to the harmful effects of AI.
• While we appreciate the nod to "environmentally friendly" applications of AI, this brief mention must be developed much further in the final guidelines. Ecological concerns must be placed on par with concerns for human wellbeing and prosperity. The final guidelines should address rights and responsibilities toward all life, ecosystems, and existing international commitments like the Paris Agreement and Sustainable Development Goals.
• The section on critical concerns, raises important unresolved questions. The challenge of covert systems is significant—whether it is identifying an autonomous vehicle on the highway, or a "bird of prey" that may in fact be a surveillance drone. This section leaves out the necessity for some of these technologies to be covert by design, especially in security and military applications.
• The section on LAWS makes no mention of the EU's direct relationship with military application of robotics and AI through the European Defence Fund. The possibility for conflict here is significant. The section makes no mention of already existing arms races, or the difficulty in dealing with illegal arms trade or negotiating arms trade treaties for AI, and the likelihood of guerilla and terrorist groups and others to subvert even the best-intentioned regulatory initiatives.
• On the point of the longer-term consequences, we consider that they go beyond the question of law and injustice, and concern the natural disposition of the human being. This is especially pertinent in artificial consciousness and moral decision-making. Religions are concerned with precisely these meta-ethical questions, and are as such an indispensable interlocutor in these conversations.

disclosure which is often incompatible with the needs and ambitions of governments and corporations.
• The section on design for all only considers the individual human and not the status or wellbeing of communities. The impact on communities and relationships is of great importance and must be reflected in such guidelines.
• The section on non-discrimination must take into account that bias and discrimination can take place long before data is collected, and technologies developed. The choosing of research agendas, and the problems we seek to solve, is an inherently biased undertaking that can only be remedied through the intentional diversification of the research field.
• The guidelines should also ask how the technologies contribute to the diversification of knowledge and how these technologies contribute directly to equity, justice, and the elimination of discrimination.
• The section on privacy should explicitly address existing applications of AI that are readily used to identify persons in public and virtual places. This includes facial recognition software and applications like speeding and toll cameras.
• The section on robustness should expanded to include more questions about how to handle attacks, security breeches, theft and illegal trade of technologies, and open source AI (and related technology). We must assume that these technologies will end up like any other commodity with problematic trade, illegal or unauthorised used, including ending up in the hands of terrorist groups. A much stronger appreciation for this is needed in this section.
• The section on human autonomy requires significant reconsideration and development. There is far more to developing "human centric" AI than simply attending to the preservation of individual human autonomy. Guidelines for Trustworthy AI must include many more questions about the impact on humans, their communities and societies, and the ecosystems that support all life on Earth. These questions could include: Does it threaten language or culture, especially vulnerable or Indigenous ones? Does the system commodify human beings and their relationships? How does the supply chain contribute or detract from the trustworthiness of the system (e.g., use of conflict minerals, slave and child labour, and so on)? How do we negotiate hybridity (virtual or actual) with the human being (e.g., brain-machine interfaces)? Does a technology or system contribute to the concentration of power in its various forms (e.g., social, political, military, and so on)? What are its effects on climate change? On ecosystems? How will this affect future generations?

dialogue. With the guidelines delivered shortly before Christmas, when it is busy for our churches and others are on holiday, it was too difficult for the Conference to consult with its membership before the dialogue seminar in early January.
Broad public consultation and debate can only serve to strengthen the spirit of these guidelines and the resultant text. The churches in particular are an excellent forum for such work as they have a longstanding interest in issues of ethics relating to all manner of technologies. This is coupled with ultimate concern for the human being, human communities and the flourishing of all life on Earth.
We look forward to the next version of these guidelines incorporating the remarks made above and those from other faith-based stakeholders.

Ashley	Williams			1. The EU's Rights Based Approach to AI Ethics (page 5) The AI HLEG rightly acknowledges that the aim of these Guidelines goes beyond Europe and aims to foster reflection and discussion on an ethical framework for AI at a global level. As such, whilst it is useful to look to fundamental rights commitments of the EU Treaties and Charter of Fundamental Rights in order to identify abstract ethical principles, we should take care to avoid producing a set of abstract ethical principles which are (or appear to be) too European centric. Likewise, we should take care to ensure these principles do not become too granular in nature. Providing a set of Ethics Guidelines that are global in application will increase the level of uptake on a global level.4. Ethical Principles in the Context of AI and Correlating Values (page 8) The AI HLEG flags tensions may arise between the principles and flags that returning to the principles of overarching values and rights protected by the EU Treaties and Charters in times of conflict. In the EU Treaties and Charters we do maintain a difference between absolute rights and relative rights. Presumably the AI HLEG envisages that the abstract principles will contain the same distinction, and this should be flagged as a means of determining how some conflicts may be resolved. 4. The Principle of Justice: "Be Fair" (page 10) The Guidelines flag the need to ensure individuals and minority groups maintain freedom from bias, stigmatisation and discrimination. Developers and implementers should be aware of the risk of calibrating AI in a way which also runs the risk of positive discrimination, which would equally be detrimental to the trustworthiness of AI. 4. The Principle of Explicability: "Operate transparently" (page 10)The Guidelines currently state that in order to ensure the principle of explicability and non-maleficence are achieved the requirement of informed consent should be sought. It is not clear here what consent the AI HLEG is referring to or in what circumstances they are suggesting it would be required. The dawn of GDPR has triggered a lot of confusion over consent and in particular when it should be required. We should take care when referencing consent in respect of AI that the same confusion does not arise. 5.2 Covert AI Systems (page 11) The Guidelines state a human always has to know if she/he is interacting with a human being or a machine; however, the action point which follows is softer that this statement – "AI developers and deployers should therefore ensure that humans are made aware of – or able to request and validate the fact that – they interact with an AI identity". Where a human is interacting with an AI and that interaction has, or may have, an impact on its decision process, then it should always be made clear at the very outset that the interaction is with AI, particularly where the interaction is with a member of a vulnerable group (such as children). A perfect example of this would be an increase in chat bots which try to mimic human interactions to increase customer engagement and influence customer behaviours.	2. Accountability Governance (page 22) The AI HLEG rightfully flags that accountability can include the appointment of a person in charge of ethics; however, we should ensure lessons are learnt from the introduction of the DPO role under the GDPR, where the attributes the DPO needs to possess and the position they need to hold in an organisation has resulted in demand outstripping the availability of expertise. This is particularly important for start-up companies who may not have the resources to make further appointments. The AI HLEG should also consider thresholds as to when it should be considered more crucial to appoint an AI ethics expert (similar to the large scale processing thresholds for a DPO).			

Anonymous	Anonymous	Anonymous	Good balance and well laid out. Scope and purpose are very clear and the drive to use Trustworthy AI as a focus to encourage development of AI across the EU and be a global leading innovator. This can be used to work with new Start-up companies to encourage to build in from the start and gain competitive advantage from the start. This is outlined in the document but could be good to mention the phrase start-ups specifically as talking about innovation. The approach of this as a set of guiding principles is the correct approach to take and building a framework that business entities can apply is very useful.	Under section 5, Critical concerns raised by AI 1. Identification without consent: The informed consent aspect will be difficult for many businesses to implement. While GDPR does cover some of this, explaining to people what the AI will do be a challenge. What level of detail does one do and how to explain. This will require careful thought. For example, data that was anonymized and a person gave informed consent to share but 5 months later a new data set is linked into the AI and it now can work out who the person is by links in the new data set to those there in place. Should one be informed as soon as anonymized data is re-personalized? 2. Covert AI systems: The need to inform from the start if it is an AI or not would not always be required, why not give it back to the person to choose. People could ask the question to that AI system if it was an automated system and it would only then inform. This way people have the choice if they care or not. If everyone is informed from the start, then people may not adopt AI and always want to talk to another human. This could stifle innovation if a person must be informed from the start. As AI grows in usage, this could be monitored and once adoption is at a certain level one could come up with a universal symbol (international standard) to identify when it is an automated response vs human. 3. Normative & Mass Citizen Scoring without consent in deviation of Fundamental rights: Would agree fully with the proposal on having opt-out options in place, and that the process, purpose and methodology for all scoring of this type is clearly explained in plain language with an option to get more detailed explanation. E.g. explain to people that their score is based on a process of getting an average across the group, and in the more detailed section it could explain the technique in more detail. This way one has full transparency as well if they wish to do a deeper dive into how it works. Challenges here with IP and Trade Secrets will need to be considered. 4. Lethal Autonomous Weapon Systems (LAWS): In this space, it is important that there is an element of human control and accountably. AI can help with more precision and target selection, but this should always remain that a human decision is required. It is important that AI is used more to supplement and not take over from Human decisions in conflicts. 5. Potential Longer-term concerns: As stated in the paper this is a highly controversial subject area. In one aspect, the principles outlined earlier cover this if one is to follow the basic principles of embedding ethical thinking into all AI practice. For the longer-term horizon and critical long-term concerns having a dedicated body to look at this could be the best approach. This could be modeled on the GDPR structure of The European Data Protection Supervisor (EDPS) where they give guidance and clarity on GDPR across the EU at a high level when new cases arise in relation to GDPR. One could have a similar body for AI which would be reviewing the framework proposed and giving guidance on how to apply the framework to the future	The 10 requirements for Trustworthy AI are a very good suggestion. Each is clear and can be adopted per situation. On the Technical and Non-Technical Methods where a request for additional technical and non-technical methods to suggest, would be happy with what is there for the first draft. As outlined in the document there are a wide array of options in this space but the ones in this document cover the main areas in a very comprehensive way.	The proposed assessment list appears to cover the key aspects from a high level and by using this a guideline would help direct questions and discussions on more specific case by case situations as they arise.

			events such as Artificial Moral agents or the possible development of artificial consciousness. It is important that the framework is something that every day business, government and institutions can follow to leverage the benefits of AI and ensuring it can put Europe at the center of the AI Trustworthy world. By extracting out the future deep thinking to a dedicated body, this would allow entities to consult with this body if they are coming close to something in this future field.		
Zora	Siebert	Heinrich-Böll-Stiftung European Union	The Draft Ethics guidelines provided by the AI HLEG are a welcomed starting point for a better understanding of Artificial Intelligence (hereafter AI) and the role of the EU in its development in a broader social context. However, while ethical commitments are a step forward in the debate about AI, they have little measurable effect on software development processes if not directly tied to structures of accountability and workplace practices. Furthermore, ethical pledges by companies or others have to be backed by enforcement, oversight and have consequences for deviation. While the suggested guidelines are in some points educational and raise awareness of particular risks of AI, ethical principles like "do no harm" or "do good" (both on page 8) are powerless if they are not backed by law. Therefore, it would have been desirable to go more into details on the current regulations in place that apply to AI and that the AI HLEG only briefly refers to (page 2).Ethical principles have been operationalised in the legal framework of the General Data Protection Regulation (GDPR) or human rights that are codified in a body of international law, entailing legal obligations and rights. They do not depend on the ethical preferences of companies. Unfortunately, the Draft AI Ethics Guidelines have little informational value on what legislation is needed; existing law is ignored and presented as part of ethics. Specific input referring to point 5. Critical concerns raised by AI• Accountability: There is an enlarging accountability gap in the use of AI. The culture of industrial and legal secrecy that prevails in AI development is a barrier for accountability in this sector. Models underlying the technology are often proprietary and systems are often untested before being deployed. Significant concerns have been raised about the lack of due process, accountability, community engagement, and auditing. Transparency has to be created at the algorithmic level, at the levels of trade secrecy laws, labour practices, and the global supply chains for example for rare earth minerals used to build consumer AI devices.• Surveillance: The use of AI can lead to amplified surveillance, especially in conjunction with facial recognition. Facial recognition threatens individual privacy and accelerates the widespread of automated surveillance. Researchers at the ACLU (American Civil Liberties Union) have demonstrated that facial recognition technology is, on average, better at detecting light-skinned people than dark-skinned people, and better at detecting		The ramifications of technological innovations like AI have to be carefully considered before it is implemented at scale (precautionary principle). There is a need for accountability and oversight in the industry. To move toward ethics only is not meeting this need. European citizens need enforceable rights that provide for accountability and redress, not lip services to ethics. In several areas those enforceable rights already exist.Recommendations:1) Take the existing legal basis into consideration and refer to it when your aim is to contribute to an informed public debate on ethical guidelines for AI. 2) Governments should execute human rights impact assessments before making use of AI.3) Foster more rigorous research on the potential human rights harms of AI.We would like to thank the European Commission for the opportunity to provide feedback on the Draft Ethics guidelines for trustworthy AI, prepared by the High-Level Expert Group on Artificial Intelligence and hope that our comments are useful.

			men than women. The risk of false positives leading to unintended negative consequences especially touches upon people of colour who are more often falsely identified. This deepens the concerns about the use of this technology and therefore the EU and its member states should critically examine and severely restrict it. There are limits of technological solutions to problems of fairness, bias, and discrimination. We have seen growing consensus that AI systems can perpetuate and amplify bias, and that computational methods are not inherently neutral and objective.• Governmental use: The increasing governmental use of automated decision making systems or the use of any form of "social scoring" or "citizen scoring" in the guise of efficiency and cost-savings directly impacts individuals and communities without established accountability structures and without adequate protections. It is a massive risk for civil rights. Example: The Federal Office for Migration and Refugees, BAMF (Bundesamt für Migration und Flüchtlinge) is using speech recognition software to protect against identity fraud. However, it is difficult to control the mechanisms of the software because we do not know on which algorithm it is based. The German government indicates an error rate of 20 percent for the software. It is irresponsible to let software decide the fate of people when we are dealing with such a high degree of inaccuracy. Other examples are the growing use of AI in the criminal justice system (especially in the USA) using risk assessment software to assist judges or predictive policing. These practises may interfere with the presumption of innocence or the right to a fair trial.• Unregulated forms of AI experimentation: New technologies are adopted quickly and tested on citizens without much regard for the impact of failures. Citizens are thus the ones to bear the burden. Example: Self-driving Uber that failed to recognise a woman and then hit and killed her in March 2018. Basic safeguards of responsibility, liability, and due process are thus increasingly urgent concerns. This example demonstrates that there has to be a plan in case something goes wrong.				
Anonymous	Anonymous	Anonymous	Thank You for addressing the challenge and the work on the ETHICS GUIDELINES FOR TRUSTWORTHY AI. It is an impressive work, a good start for fostering a high quality discussion.	The impact of the AI Systems Operators on the Labour Market and work situation of weak social groups is neither analyzed nor addressed in the discussed Document The technological revolutions linked to coal and steam engine, rail transport, automobiles, long distance media, computers and mass communication were bound by the physicial propagation of the disruptive technologies. It means that there was enough time for people working with old technologies to readjust their careers. On the other hand, the propagation of the AI Systems is fast, unrestricted and uncontrolled. The phase-transition linked to the usage of AI Systems is comming and is unavoidable, but the change and risks associated with the change should be under control. In the workforce situation and job market chances of weak social groups is at a huge risk which is neither understood or	The Safety First Principle is needed as in case of all real technologies on which life or health of human beings are dependent The discussed Guidelines document is a good starting point for fostering a structured discussion. Nonetheless we should keep in mind words of Prof Richard Feynman who concluded his presentation of findings from the Rogers Commission report (after the catastrophe of the Space Shuttle Challenger in 1986): "For a successful technology reality must take precedence over public relations, for nature cannot be fooled.". In this context, it should be made clear that the discussed framwork unintentionally promotes uncontrolled risk taking and indirectly leads to information concealment. There essentially three key components which usually lead to man made catastrophes: - there is single point of failure embedded in the uncontrolled, non-transparent data	The missing rule "skin in the game" i.e. the fudamental rule of managing the underdefined complex risks The focus of the discussed document is on the technology and not on the AI Systems Operators. The proposed Guidelines don't address the problem of complex underdefined risks embedded in the AI Systems because there is no serious approach to making AI Systems Operators accountable, responsible and liable for the risks embeded in technologies they capitalize on. As we have wittnesed from the recent Credit Crisis (2008) and Greek Soverign Debt Crisis (2011), lack of responsibility, accountability and liability amongst the market (economy) participants who used advanced technologies (including AI algorithms for risk management and trading) led to huge disasters with impact on lives of EU. citizens In case of the both catastrophes the three	Disclaimer: the presented opinions are solely my own and not the views of any of my previous employers or of a current employer.

addressed by the politicians. To make things maybe more realistic for the employees of the European Commission: applying the currently available AI technologies and keeping the current level of the quality of outcomes at the European Commission, the budget of the European Commission can be cut by 20%, the administrative and clerical workforce cut by 40% and taking into account those two cuts it would not influence and in many cased increase the quality of work at the EU.

Although as B. R. Ambedkar had noted: "History shows that where ethics and economics come in conflict, victory is always with economics. Vested interests have never been known to have willingly divested themselves unless there was sufficient force to compel them.", the discussed Guidelines document shows that there still seem to be some political willingness to properly address the impact of the AI Systems Operators on lives of the EU citizens. In my opinion, we should include in the started discussion about the AI Trustworthy Systems also aspects of Risk Management, Governance and Compliance of AI Systems Operators within local labour markets.

processing, model design and mass-deployment carried out by AI Systems Operators
- lack of full personal liability for failures of the management board members at AI Systems Operators
- lack of unconstrained financial liability of the AI Systems Operators

Other types of technologies like in pharma, transport, contruction or banking are under much stricter rules and the prerequists for technology operators are: responsibility, accountability and liability. In this context, the AI Systems should be treated as a technology and not a mystique being. There is no reason to approach the AI technologies in a less scientific and organized way than to technologies from other industries like pharmaceutical technologies, medical equipment, nuclear technologies, aircraft or aerospace. Much more rigorous approach is need in defining the Trustworthy AI systems because it is a technology, maybe revolutionary, but only technology and should be treated as such i.e. by tailoring and applying existing technical standards applicable in the safety critical environments or working on new standards for the AI systems from the ground up. The EU has a long history of developing safety standards and this many years experience and know-how should be capitalized.

problems were present :
- potentially skewed data and non properly managed models are the source of an exploding problem
- there is a lack of proper structure for personal liability and financial liability for decisions about using and operating complex systems and structures
- there is a lack of proper compensation for losses incured by the systemicaly important corporations

In case of the AI Systems Operators we can talk about a similar situation: a few of huge technology players (also called "too big too fail") leverage their cutting-edge technologies and big data resources to provide innovative AI based services but without taking full responsibility, liability and accountability for the risks. The mentioned players had already shown how they treat Privacy, Safety of Personal Data and Local Economies.

So it is not the technology itself or a developer of technology be held responsible, accountable and liable but an Operator of the AI Systems. The discussion about the ETHICS GUIDELINES FOR TRUSTWORTHY AI should be reframed.

Louis-Marie Bès

Three main comments to the “Draft AI Ethics Guidelines for Trustworthy AI”Considering the High-Level Expert Group on Artificial Intelligence as mainly composed by representatives from industry and federation, the possibility given by the European Commission to take part into the drafting of the guidelines is really welcomed. Those should be shaped in order to guarantee the necessary respect of human rights and dignity by Artificial Intelligence and its implications. As a preliminary remark, considering the seriousness of the challenges caused by Artificial intelligence, a non-binding and voluntary set of principles would be insufficient and would fail utterly in its purpose. In this regard, the upcoming guidelines should be mandatory and should apply to all stakeholders, not only to signatories. If not, companies may consider ethical values and principles as burdens hindering innovation and profitability and may brush them aside. In that respect, it is fundamental that companies and key actors be as far involved as possible and not be tempted to limit their commitment to a misleading “ethical-washing”.Goal of AI ethicsHaving regard the vision introduced on page 2 (“The goal of AI ethics is to identify how AI can advance or raise concerns to the good life of individuals, whether this be in terms of quality of life, mental autonomy or freedom to live in a democratic society”), it seems necessary to propose a new definition, or at least to complete the existing one. Introducing AI ethics as only the way to extend its benefits to all and not only to a minority of winners seems incomplete. Thus, the goal of AI ethics should be defined as follows: “it aims at identifying how AI can advance or raise concerns to the primacy of humankind over machine and robotics, to engrave the pressing need to guarantee human rights

over technology, to protect the human capacity to choose and freely interact, evolve and make decision autonomously. The goal of AI ethics is to enshrine the principles of autonomy and informed consent. It concerns itself with issues of diversity and inclusion (with regards to training data and the ends to which AI serves) as well as issues of distributive justice (who will benefit from AI and who will not)".Human responsibility of algorithmsThe Joint Research Center report "Artificial Intelligence: A European Perspective JRC" calls for a cautious approach regarding interaction of humans and machines in a more and more technological progress-based world. On the back of this observation, it is highly unthinkable that human rights and responsibilities be diminished due to an increasing interaction between humans and machines. This would be the very and major comment: accountability is ultimately related to human responsibility and the creation of a specific legal status for robots is a Promethean idea, which must be forgotten. Human beings cannot be unloaded from the burden of responsibility. Human enhancement technologies and transhumanismLast, the AI ethics guidelines should set strong red lines regarding augmentation of humans thanks to robotic. To this end, the High Level Expert Group should define, in the sharpest way, what is restorative care and what is not.A limitless understanding of augmenting humankind would lead to an unfair and unequal society. "All men are created equal" and we should keep it that way.********

Anonymous	Anonymous	Anonymous	Good one!	Concerning "2. From Fundamental rights to Principles and Values"In my opinion informed consent is one way to ensure that human choices are indeed free. But seeing how careless lots of people are giving their consent without consideration nowadays especially in relation with digital media etc., I believe this goal is very hard to achieve. When I read this part the following comparable situation came into my mind: I think a lot of people especially when they're buying some stuff via internet they tend to accept the general business terms without thinking about the consequences. And why is that? Well, the easy answer to this is, everybody is doing is this way, no one actually reads them. Does this practice arise further questions concerning the freedom of human choices? No. Thanks to the statutory control of general terms and conditions set out in the COUNCIL DIRECTIVE 93/13/EEC on unfair terms in consumer contracts, the careless acceptance has a somewhat limited range. I'm not sure whether something like this can be adopted for the AI issue, but I think there will always be some people out there who never reach the point of being informed in order to give their informed consent. And they need our help as well, maybe more than others. Failing in helping them will only create a breeding ground for extremists against AI in general.Concerning "3. Fundamental Rights of Human Beings"I agree that it is absolutely crucial to somehow teach an AI "that all people are treated with respect due to them as individuals, rather than merely as data subjects." This is what humanity is all about, being recognized an individual no matter what.I also want to emphasise that "citizens should enjoy a right to be informed of any automated treatment of their data by government bodies, and systematically be offered to express opt out." Tough I tend to not only include government bodies in here, but also to companies who have similar power as states, some may say so called corporate nation-states.Concerning "4. Ethical Principles in the Context of AI and Correlating Values - The Principle of Non maleficence: "Do no Harm"In my honest opinion a priority should be "to develop and implement AI systems in a way that protects societies from ideological polarization and algorithmic determinism." I'm saying this because of the huge impact algorithms nowadays already have especially in terms of gathering news via social media. Seeing that real life conversations take place less and less, but instead more and more conversations using digital methods, algorithms "decide" what kind of news is shown to you. Making you see only the things you want to see, like you're living your nice little life in a bubble. But it is essential for a society to interact also with people who have different opinions, otherwise the world we're currently living in is slowly falling apart.Concerning "4. Ethical Principles in the Context of AI and Correlating Values - The Principle of Justice: "Be Fair"In the draft it is written that "Justice also means that AI systems must provide users with effective redress if harm occurs, or effective remedy if data practices are no longer aligned with human beings' individual or collective preferences." Provided that the final guidelines are indeed just non-legally binding guidelines I think, that this goal is hard to achieve. Why would any company or	(Sadly, I didn't find the time to read chapter II)	(Sadly, I didn't find the time to read chapter III)	Generally speaking algorithms are primarily satistics orientated and thus aim by design for standardization. Along with this comes the negative effect of deindividualization. The indiviuality of every human being becomes less and less important. Maybe we won't have one algorithm for a specific task in the future but lots of different personalized algorithms?Another main point in this discussion is in my opinion how do we achieve the beneficence of algorithms. Historically speaking the power always belongs to the state, but nowadays the most powerful algorithms are controlled by private companies. A slowly shift in the balance of power. How do we want to design it?

its managing director voluntary acknowledge a failure within its AI in relation to the human-centric approach? The decision making in a company is primarily determined by money, money and money (in its various expressions, e.g. its reputation).Concerning "5. Critical concerns raised by AI -5.2 Covert AI systems"Another very important point are covert AI systems. To my understanding these covert AI systems already have a huge influence nowadays. I'm thinking of social media bots, as far as I know they're getting better every day and thus it's not so easy to detect them. For normal users they appear to be just like them – normal. But instead it's just an AI identity, waiting to influence you. I can only speak for myself, that I can only determine the badly programmed AI identities Concerning "5. Critical concerns raised by AI – 5.3 Normative & Mass Citizen Scoring without consent in deviation of Fundamental Rights"I believe that the today existing domain-specific scoring is in no way comparable to the normative & mass citizen scoring. The examples given either concern basic education or voluntary stuff, and besides that the scoring factors are all transparent and everyone is able to look them up with a little effort. I for myself don't want to live in a world where everything is determined by your score, like it is shown in the Black mirror episode 'Nosedive' https://en.wikipedia.org/wiki/Nosedive_(Black_Mirror).Concerning "5. Critical concerns raised by AI – 5.4 Lethal Autonomous Weapon Systems (LAWS)"Ideological speaking peace is always the best, but the world is consistently on fire. So I think that an arms race relating to LAWS in inevitable as not everybody, not every country is sharing the same core values as explained above. Just have a look at some UN decision, it's kind of ridiculous how many nations there have a different understanding of human dignity. Thus I feel way more confident when I know that states who share the above describe core values have access to these technologies. In the best scenario I hope for a somehow 2nd cold war situation, lots of options but no real war as the risks for either side are just too high. Having said this, peace for everyone!Concerning "5. Critical concerns raised by AI – 5.5 Potential longer-term concerns"Often long-term developments are foreseen and illustrated in movies. Almost everybody knows the movies with Arnold Schwarzenegger as Terminator, where the human race is desperately trying to trigger the kill switch. But there might also be a possible future like it is shown in the movie Transcendence (https://en.wikipedia.org/wiki/Transcendence_(2014_film), in which it is not so "easy" to find/develop a kill switch. I have no clue how to face such a possible arising problem, though a risk-assessment approach seems reasonable. In this regard it might be helpful to follow the Telekom AI guideline no. 7 "We are able to deactivate and stop AI systems at any time (kill switch). Additionally, we remove inappropriate data to avoid bias. We have an eye on the decisions made and the information fed to the system in order to enhance decision quality. We take responsibility for a diverse and appropriate data input. In case of inconsistencies, we rather stop the AI system than pursue with potentially manipulated data. We are also able to "reset" our AI systems in order to remove false or biased data. By this, we

				install a lever to reduce (unintended) unsuitable decisions or actions to a minimum."			
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Erlin	Gulbenkoglu	Silo.AI	This part of the document tells us the motivation of the EC and AI HLEG. We are wondering if the motivation to choose the word “trustworthy” could be explained. We mean that the need for humans to “trust” AI could be explained more in detail for readers to comprehend why this adjective is chosen to frame the discussion.It was clearly pointed out that ethical principles should be operationalised by integrating them into the design, development, implementation and use of AI. We still think that there could be more attention given to the design phase of AI systems which is the step where one needs to define the purpose of the system, the necessary data, methods to ,, architecture of the system, privacy & security procedures in order to have a preventive and proactive system.	We appreciate the association of 5 principles with correlated values. We believe the phrases for values are also catchy (e.g. “Do Good”, “Do no Harm” etc.).We feel the description of each principle has too many components and has wide coverage. The content for each principle could be more organised.AI is only as unbiased as its developers are, which is why there should be guidance and incentives to get people from different backgrounds, educations, ethnic and other groups to develop and test AI. It is not just about vulnerable demographics but demographics that currently do not participate in the development of AI. This could be stressed more among the ethical principles, perhaps under “Do no harm” value. We believe that this approach will naturally help different demographics to be taken into account with greater attention. We would here need clearer and more concrete longer-term ways to ensure that different groups of people are both developing AI and hence taken into account. Another aspect reflecting the European values would be considering our diverse languages. We should make sure that both developing and the tools that use AI include the European languages. European people need to be able to use AI with their own language to make sure there is no gap	Perhaps the Requirements of Trustworthy AI and Technical and Non-Technical Methods to achieve Trustworthy AI sections could be merged somehow since there are intersections of the topics mentioned.Regarding the trade-offs for transparency or correcting bias, I think many of these could be documented even at an EU level on a general catalogue of trade-offs likely to happen. This way, everyone wouldn’t have to come up with their own version of the same issues.We believe the technical methods mentioned are explained at a high level which, we think, is even difficult to concretise and see the actually mentioned practices for a technical person. We think it would help the reader to see concrete practices for each example in the technical method section.Furthermore, Privacy by Design principle was mentioned in the GDPR as well without any practical details. Therefore, we were hoping that this guideline would say something more practical and concrete about how to incorporate this concept into the design of the AI systems.Regarding the Explanation (XAI research) in particular, we think it is important to stress that with the new techniques introduced in XAI research we are also able to provide the individual level of explanations to ML model predictions	We believe introducing practical questions for each requirement mentioned in the previous section is very practical and beneficial for organisations to manage to follow the requirements. Perhaps, it would be easier to follow the questions if they were presented with the requirements.Some suggestions for the questions would be:Data Governance1) How are the data transfer and erasure processes governed?2) How do you assess the scope of data and define what data is necessary before designing the AI system? (Assessing the purpose of AI and then defining the scope could be practical)Governing AI autonomy1) What specific ways does the AI system have a human-in-the-loop? (e.g. data assessment, labeling the data, training the model, feedback loop, communicating AI’s predictions to a user (model explanation), continuous improvement of models through feedback)2) Is there continuous verification of fair and unbiased operation of the AI system?Respect for Privacy1) Is there any data anonymisation (or pseudonymisation) technique in place to ensure data minimisation, meaning that no unnecessary data will be seen/used?2) Is there any threat to linkability attacks (e.g. if you have just pseudonymised some fields of the data but the rest can be still linked to another dataset	We would like to thank the European Commission and the HLEG for such an effort. We find the guideline contentful and inclusive about many discussions we have encountered. But, this brings the burden of difficulty to see the big picture.We had difficulty to understand the structure and the analytical order of sections of the document. We suggest a reorganisation of the sections and subsections based on a more thorough classification of the content. For example, splitting the document into more chapters so that each chapter has a particular focus. We believe forming a more analytical structure will make it easier to understand the document.Even though there are many different interests that might affect the making of this document, we believe EU deserves and needs clear and well-structured ethics guidelines for the development of AI. With a well-organised document Europe will be better prepared to lead the way for safe and human-centric AI.Note: This review was written by Erlin Gulbenkoglu & Pauliina Alanen & Mark van Heeswijk

			between people that are able to use AI in English and people that would like to use them in their mother tongue. We believe Section "5. Critical concerns raised by AI" could be a separate section at the end of the document or in the appendix. We believe that section includes more concrete cases for concerns which is hard to associate with the more high-level narrative of the chapter.	(local explanations). Which is very important for the transparency of each prediction made about an individual. We believe it is also crucial to require user interfaces which help users to easily comprehend the explanations provided by the XAI methods.	to reveal the identities of individuals)?Safety1) Are users knowledgeable about the security features and best practices to manage the security of the system?Transparency1) Are the explanations provided to users/stakeholders understandable irrespective of people's education or comprehension level?	
Sue	Daley	techUK	techUK welcome the opportunity to provide feedback on the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) Draft Ethics Guidelines for Trustworthy AI. Based on the points raised in this response techUK would welcome further detail, clarity and consultation on the concerns being raised below. For the final guidelines to be well-received by the AI community, the guidelines must be robust, credible and reflective of the amendments made during the consultation process. techUK would strongly advocate for a second consultation on the amended guidelines before the final version is published. We particularly foresee that additional work and consultation will be required on Chapters 2 (Realising Trustworthy AI) and 3 (Assessing Trustworthy AI). Executive Summary We support the AI HLEG's ambition to produce a set of AI Ethics Guidelines for Trustworthy AI that can be commonly adopted across the European Union. We advocate the concept of these guidelines being a flexible, evolving and a voluntary tool kit for industry and other stakeholders developing and deploying AI, now or in the future. We also support the decision to produce guidelines rather than taking the regulatory route, which would risk stifling innovation. Furthermore, it is worth acknowledging that many companies are already doing a significant amount of work to ensure that ethical practices are embedded throughout their organisations. Where these guidelines can add real value is in the form of operable guidance, building on the AI principles that we've seen being developed at a company-level. GlossarytechUK support that "Trustworthy AI" has two components: an "ethical purpose" and the need to be technically robust. This will help to ensure that AI is trusted by its users and in turn improves Europe's competitiveness in this market. However, we would recommend that the AI HLEG uses the term 'ethical intent' instead of 'ethical purpose' in the amended version of the guidelines. Although it is important that designers of AI systems are clear on the purpose of a system in order to build user trust and good governance. Our view is that the current terminology could be confusing for both designers and consumers if the	There is concern that in many of the descriptions of AI in this chapter the tone is often negative towards AI technology. This section should be amended further to reflect the positive potential and opportunity for AI to be a power for good for individuals as well as society as a whole. 3. Fundamental Rights of Human Beings In this section of the guidelines techUK would like to see further discussion on a number of the statements made. For example, under "freedom of the individual (3.2)", the guidelines could offer more real advice for organisations on how to address concerns relating to striking the right balance between the freedom of an individual and that of society in specific contexts or circumstances where conflicts between the two might occur. Further clarity would be welcomed on how this section also fits with the national security obligations/requirements of government. Also under the title of "citizen rights (3.5)" while under most circumstances we agreed that citizens should 'systematically be offered to express opt out', the guidelines fail to acknowledge that there may be a limited number of specific circumstances in which an individual's decision to opt-out should not apply, for example when there is an overriding public interest, such as the monitoring and control of important diseases in humans such as TB and diseases of epidemic potential such as Ebola. Similarly, the right to opt-out of criminal investigation systems might not be appropriate. This point should be reflected in the final guidelines. 4. Ethical Principles in the Context of AI and Correlating Values techUK sees the five 'principles' highlighted in the document: do good, do no harm, preserve human agency, be fair and operate transparently, as sound. However, given the issues each principle covers there is potential for overlap with many of these principles when they are operationalised. It is therefore not clear how the principles would work together when put into action. It is suggested that the guidelines offer advice on how to handle such situations. The principle of 'Preserve Human Agency' is seen as particularly critical from a legal and practical point of view as it mitigates the risk of system-determined decisions not being checked. However, the guidelines also state that a consumer or user	While the executive summary of the draft makes it clear that the document offers guidelines and a framework for stakeholders to follow to achieve trustworthy AI, there is concern with the terminology used in the following section of the document. Particularly as it focuses more on requirements rather than advice. We would recommend using softer terms like 'recommendations' and 'advice' in this section to reflect the voluntary nature of the guidelines. In addition to the requirements that are already discussed, techUK believe demonstrating why we are developing AI systems and showcasing the positive impact that this technology is having on peoples' lives is key to building public trust in AI. It is suggested that the importance of engagement with users, as and where appropriate, should be seen as a key requirement for building trustworthy AI. Engaging at the citizen-level is also key to prevent the AI ethics community from becoming insular or disconnected for the societal changes happening all around us. This should be reflected in this section of the guidelines. 1. Accountability techUK felt that this paragraph on accountability focused too much on addressing issues of liability and monetary compensations in the event of an incident. It does not offer guidance for organisations to consider the steps and measures that may need to be put in place to ensure there are appropriate policies and procedures for determining where accountability and responsibility rests for the different stages of the design, development, adoption and deployment of AI systems. It's also insufficiently clear how this current definition of accountability will intersect with existing legal requirements. 2. Data GovernanceGiven the existence of established data governance frameworks and best practices it is suggested that this section highlights what already exists in this area that could be directly applicable to AI ethics. For example, the requirement within GDPR should be highlighted. The section as it is currently drafted does not mention the security of data (in use and at rest) which along with availability, integrity and usability is a key component of data governance. Further clarity is also sought on what is meant by the term "high quality AI" and how	While the third section starts to helpfully sets out some practical steps and questions to ask to operationalise the guidelines, further clarity and detail is needed for this objective to be fully achieved. For example, currently the questions lack a level of detail and in some sections (such as Respect for Privacy) many questions appear to repeat themselves or overlap. We also believe further clarity is needed on who the current assessment list of questions is aimed at specifically. The document states that the assessment list is for "AI developers, deployers and users to operationalise Trustworthy AI" but it is currently unclear which stakeholder groups should be considering, and answering, each of the questions. As they are currently drafted, we would anticipate that these questions should be asked by those developing the AI systems rather than the users of AI systems, however this isn't clear. If the responsibility is placed on developers, this document doesn't recognise the responsibility of the wider ecosystem, for example the buyers, suppliers, etc.Also, it is seen as unfortunate that the questions do not currently mention, or link to, existing regulatory frameworks, particularly as GDPR, which could provide further detail and advice to those answering the questions. To help with the practical, operationalisation of the assessment list, the HLEG have committed to introducing four use cases in the final version. These case studies will be instrumental as to how the guidance is interpreted, therefore techUK would recommend that these use cases are reviewed as part of a second consultation. The HLEG have suggested that the use cases should represent four sectors- Healthcare diagnosis and treatment, autonomous driving/moving, insurance premiums and profiling and law enforcement. We'd encourage this list to be extended to represent a wider range of sectors. In April, the ISO/IEC JTC 1/SC 42 are launching a global use case inventory. It would be helpful, where possible, to align these two initiatives. There are key issues that must be addressed before moving forward and where further consultation is needed before the guidelines are finalised.	Accessibility of the guidelines For many companies, particularly AI start-ups and SMEs looking to adopt and deploy AI, there is real concern that the level of granularity in the guidance could discourage their use. A number of our members have flagged that the guidelines are difficult to navigate and that they would struggle to operationalise the document in its current form. There is a need for the guidance to strike a balance between clarity and interpretability, whilst retaining credibility. Tone and terminologyAs mentioned above there is concern with the tone of the draft given that the aim of the document is to offer guidance and a voluntary framework to be followed. The start of the guidelines recognises that different contexts will require different approaches, with flexibility required in application. However, this principle could be better reflected in the guidelines which frequently refers to terms like 'requirements or 'compliance.' We would recommend using softer terms like 'recommendations' to provide a degree of flexibility instead of suggesting hard rules.It is also not clear where the responsibility lies for the ongoing long-term operationalisation and monitoring of the guidance once they are finalised. This is an area we suggest the AI HLEG considers particularly given the number of initiatives happening in this area across individual Member States. We're also interested to understand how the AI HLEG see these guidelines working in relation to other Member State, European and International context guidance that currently exists or may be developed in the future.

stated purpose of every AI system was expressed in terms of ethics. Rather than combining 'purpose' and 'ethics' we would suggest that those developing AI have a clear statement of purpose, for example outlined in their company principles, and that a separate assurance should be provided around the intention to design an AI system in line with the AI HLEG ethical guidelines. Introduction: Rationale and Foresight of the Guidelines techUK agrees and supports the AI HLEG's acknowledgment of the need for a human-centric approach to AI. However, with regards to the proposed sign up process for the guidelines further clarity and consultation is required on what the endorsement process would will look like and involves before moving forward. It is not clear how an endorsement process that requires organisations to sign up to the guidelines fits with the voluntary nature of the guidelines. Careful consideration is needed to ensure that these principles are adopted across and beyond Europe. It may be useful to learn from similar initiatives that have been adopted successfully, such as the Online Child Protection initiative.

of an AI system has 'a right to opt out and a right of withdrawal'. techUK would argue that there may be specific circumstances, for example national security, where a general right to opt-out or withdraw could be detrimental to others. As the GDPR already covers rights in this area, particularly where automated decision making is taking place, it is suggested that guidance in this area follows the requirements under GDPR only. While techUK supports the principle to "Operate Transparently" the text in this section is seen as too simplistic in its current form. As written this section would be considerably challenging for those developing deep learning systems with complex neural networks. Also given the requirements outlined in this section there is a risk that the guidelines could be setting unrealistic expectations that could hold back and stifle AI innovation in Europe. We would suggest introducing a line on the need to advise those developing AI systems to continue to improve measures taken to better explain the decisions made by deep learning systems. The requirement for "business model transparency" particularly on the "intentions of developers" in this section is also seen as a step too far and could also stifle AI innovators coming to Europe. While transparency is important to developing trust and confidence in AI systems, the guidelines should focus on how to ensure that the right mechanisms are in place so that the decisions and outcomes made by AI systems are transparent, fully understandable and open to challenge and redress by both businesses and users. 5. Critical concerns raised by AI techUK recognise that a number of the 'critical concerns raised by AI' need further thought and discussion. However, from techUK's perspective we think that the addition of section 5 currently risks clouding the discussion, causing possible confusion and drawing attention away from the purpose of the guidelines, which is to help organisations developing and deploying AI. Many of the statements and issues raised in this section relate not only to AI but to other technologies that are not examined in more detail in the guidelines. Also, it is not clear how statements relating to androids being "covert AI systems" and robots being "built to be as human-like as possible" are helpful to organisations that want to embed ethics into their business processes and do the right thing. The issues raised in this section are complex and cannot be debated in silos. techUK would recommend a series of face to face round table events to explore and discuss the nuances of these debates before including them in the final guidelines.

this would be defined. The first paragraph of this section recognises that 'databases gathered inevitably contain biases.' and the need to "prune" bias out of data. While it should be recognised that there will be underlying bias in training databases available, this statement could be misleading as it may never be possible to completely remove bias from datasets. Instead we would propose that the guidelines should encourage those creating AI systems to also ask - Are potential biases in the data examined, well-understood and documented and is there a plan to mitigate against them? All the guidelines should stress the responsibility of those developing AI systems to ensure they use good quality training data.3. Design for allWhilst the concept of designing systems that allows all citizens to use the products of services is laudable, the guidelines should recognise that different AI systems will be developed for different users and purposes. For example, this could include AI systems that don't directly impact on individuals such as internal efficiencies and supply system operations. In addition, it is likely that some AI systems will be developed targeted at particular groups, for instance because of the availability of suitable training databases. What is key is to recognise where there are limitations, and where it is practicable, to seek to expand their applicability and accuracy to wider populations. Putting in place a blanket ban on non-universal systems is likely to severely limit the development of new AI applications. The guidelines should therefore take the wider development and use of AI into consideration.4. Governance of AI Autonomy (Human oversight)The final paragraph proposes that "the user of an AI system, particularly in a work or decision-making environment, is allowed to deviate from a path or decision chosen or recommended by the AI system." techUK would suggest that there may be certain circumstances, for example cybersecurity, where allowing any user of an AI system to deviate away from a choice made by an AI system could be detrimental. This statement should be amended to reflect that in such situations a person with the appropriate level of authority and qualification, for example a cyber security expert, may be able to deviate from decisions made by AI systems. This statement should be amended to reflect this point and also highlight the potential role for risk management which is lacking from the current draft. 5. Non-discriminationThis is clearly an important area, and techUK supports the ideas in this paragraph. It is also worth noting in the guidance that there might be circumstances where positive discrimination in AI algorithms might make sense, for example to compensate societal biases such as gender stereotypes for particular job roles. Instead the guidance should reflect the need to ensure that bias does not lead to unfair discrimination. 6. Respect for (& Enhancement of) Human AutonomyWhile the concepts explored in this opening paragraph of this section is supported, the remainder of this section is unclear as to the guidance being provided. The use of terms such as "extreme personalisation" and "nudging" without any further clarification or detail is seen as unhelpful in a document that is aimed at providing advice and guidance. The tone of this section is seen as negatively

focused and does not reflect, or offer any examples, of how AI systems could be used to enhance human autonomy. We would welcome further input and explanation to this section before the guidelines are finalised. 7. Respect for privacy In what is a surprisingly short paragraph, techUK would like to see the importance of GDPR outlined more clearly at the start of the section. Also, this section could benefit from introducing advice and guidance on how organisations might deal with concerns related to the balance between human rights and ethics and the relationship between privacy and security. Similarly, there's no advice on issues such as the relationship between privacy and the use of inferred data.8. RobustnessIn the section on "Reliability and Reproducibility" while the guidelines touch on the difficulties AI systems currently face to "reproduce results", techUK would like the guidelines to reflect in more detail the nascency of AI technologies and the fact that AI systems continually learn and improve over time. From a practical point of view this currently means it may make it difficult to guarantee reproducibility. The guidelines should therefore reflect the current stage of development of AI systems and update the guidelines accordingly as AI evolves. 9. SafetyIn this section techUK would like to suggest that guidance is provided on the need for clear information on where responsibility for safety of AI systems sits within an organisation. Also, further detail is also needed on what "formal mechanisms" to measure the "adaptability" of AI systems are being envisaged. Related to this section is a need for clarity on many of the terms used in the assessment list on safety found on page 27. For example, it is unclear what would be considered a risk to "human physical integrity". 10. TransparencyTransparency is key to building trust in AI systems. However, a requirement for those developing AI systems to be open about "development processes" could put European AI companies at a competitive disadvantage in what is a global marketplace. It is suggested that the guidelines should focus more on the importance of openness on how decisions and outcomes are made by AI systems and how these decisions can be challenged. The current guidelines also do not consider where different levels of transparency are necessary. It is important that there isn't a blanket prescription for all AI systems. For example, in the case of using AI to prevent and detect financial crime, individuals should not be aware of how the technology works, as it would risk undermining the purpose of anti-fraud detection systems. However, the system should still be able to explain why a transaction was identified as fraudulent and how that decision could be challenged. Transparency is also important in the identification of problems with AI systems. There needs to be a transparent governance process for identifying and reporting failings in systems and getting faults corrected.Technical and Non-Technical Methods to achieve Trustworthy AITechnical MethodsIt is suggested that this section could be improved by adding further detail and explanations to a number of terms being used. For example, there is a lack of explanation and detail on such terms as "values by design", "sense-plan-act". For these guidelines to be understood and used

by a broad range of stakeholders the additional of further detail to this section would be considered useful. • Traceability and AuditabilityWhile we support the importance of audits the guidelines should explain that auditing of AI systems could involve auditing of software use to create and run the AI system as well as the computer models being used and the data used to create these models. Also there is concern that the guidelines do not acknowledge the importance of putting in place appropriate safeguards that organisations may need to consider to protect commercially confidential information when audits are conducted. This should be included in the guidelines. Non-Technical Methods • RegulationThe first paragraph provides examples of the regulations that exist today to increase AI's trustworthiness. While it mentions "safety legislation and liability frameworks" it fails to mention data privacy law particularly the GDPR. Given its role in increasing transparency and trust at the societal-level, GDPR should be viewed as the legal baseline of an ethical approach to technological adoption and included in this section. techUK agree with the AI HLEG that Trustworthy AI requires responsibility mechanisms that, when harm does occur, ensure an appropriate remedy can be put in place and welcomes inclusion of this statement in the guidelines. • StandardstechUK would suggest using existing standards mechanisms, which already incorporate the views of stakeholder communities in Artificial Intelligence, to take the guidelines forward. We'd recommend that the AI HLEG aligns with existing initiatives such as the work of the CEN-CENELEC Focus Group on Artificial Intelligence and the ISO/IEC JTC 1/SC 42. • Accountability GovernancetechUK believe the guidelines should encourage the importance of organisations having both an internal ethics panel or board and an external ethical panel or board. Conversations around ethics should not be happening in silos and it's important that businesses engage in these issues internally with employees and externally with stakeholders. However, the guidelines should also reflect that for many organisations, particularly AI start-ups and SMEs, this may not always be feasible. The guidelines should therefore provide additional advice and guidance on existing European or International initiatives that could help them organisations to ensure appropriate measures are in place. For example, the guidelines could include a list of initiatives at a Member State level in this area that could provide more advice. For example, in the UK the Centre for Data Ethics and Innovation, the Information Commissioner's Office (ICO) and the Ada Lovelace Institute. • Education and awareness to foster an ethical mind-setThe guidelines should offer additional guidance to organisations on how to ensure appropriate and suitable training is put in place that reflects the capabilities, limitations and adaptability of the AI solutions being used and any governance mechanisms that may be required in this area. • Stakeholder and social dialogueStakeholder and social dialogue is fundamentally important if we wish to move the ethics debate forward. If society cannot trust and have confidence in AI there is a risk that the potential of AI will not be full realised. Further thought and

					direction is needed on how this work happens in practice. It will most likely require a lot of co-ordination following the publication of the final guidelines and it is not yet clear who will ultimately take the lead going forward. The important role of stakeholder engagement to assist in the societal acceptance of AI technologies should also be acknowledged and addressed in the guidelines, it may be useful to learn from other complex developments that have generally been successful accepted by society. For example, the establishment of the Human Fertilization and Embryology Authority (HFEA) in the UK which played an instrumental role in taking a trusted position on behalf of society around the most difficult ethical issues surrounding embryology and associated genetic manipulation. Similarly, we'd encourage the AI HLEG to look at how the aeronautics safety systems industry have addressed safety and ethics issues.• Diversity and inclusive design teams Ensuring diverse and inclusive design and development teams is essential for building AI systems that people can trust. Organisations developing AI need to consider this as a criterion from the inception of an AI system right up until it's deployment, and beyond.
Yuanyuan	Xiao	BTC Business Technology Consulting AG			Authors: Yuanyuan Xiao, Till Luhmann, Michael Stadler, BTC Business Technology Consulting AGOur comments pertain to Section II.1 – Requirements of Trustworthy AIWe think, that requirements should be differentiated along several dimensions: Their relevance for public acceptance and their relevance to specific types of AI-systems.The benefit of this more differentiated approach to requirements is a higher acceptance of requirements by developers and a more specific discussion of AI-systems by potential users and the public in general.First, we classified requirement based on the costs incurred for their technical realization. This is an important factor when judging the possibility of implementing the given requirements.Second, we propose to classify the given requirements by their relevance for public acceptance of AI. If the requirements most relevant for public acceptance are given a higher priority when discussing and implementing AI-Systems, this will result in a higher overall acceptance of those systems.Third, we propose to introduce a matrix allowing to differentiate the relevance of requirements along different categories of AI-usage, the categories being decision-supporting AI, decision-making AI and autonomous AI. By decision-supporting AI we mean AI-systems that support decisions of humans towards a well-specified goal. The decisions are always and consciously taken by humans after reviewing

the proposals made by the AI. Decision-making AI denotes AI-systems that routinely make decisions, which however can be overruled by humans. The term autonomous AI refers to AI-systems that provide no means of interaction by humans. That means, the AI decisions cannot be overruled by humans in normal operation of the system apart from switching off the AI (and thereby rendering useless the technical system controlled by the AI).Essentially, not every requirement is relevant for every type of AI-system. We therefore propose, only to apply the requirements classified as having a high or very high relevance for a given type of AI-system to that type, while requirements classified as having a low relevance for that type of AI-system do not have to be applied to it. For example, the requirement “Governance of AI Autonomy (Human oversight)” is highly important for decision-making AI-systems and extremely important for autonomous AI-systems but not so important for decision-supporting AI-systems.A proposed classification and an assessment of requirements for the mentioned categories is given in a table which we sent to CNECT-HLG-AI@ec.europa.eu via e-mail.When analysing the requirements proposed by the paper for evaluation of relevance and for classification, our discussion yielded some additional comments:1. Requirement #1 “Accountability”: A measure and a definition of output quality, related to individual use-cases should be defined. Thereby, AI-quality could be defined as deviation between actual output and output promised to users based on quality of the data, the AI operates on.2. Requirement #2 “Data Governance” describes a technical foundation. It is important to minimize the possibility of data manipulation.3. Requirement #3 “Design for all” should not be applied to AI-systems in general, since each system is developed for specific purposes having different user groups. It should be specified in the paper, that this requirement is only relevant for systems designed for public use. It is not necessary to follow this requirement for each specialized system.4. Requirement #4 “Governance of AI Autonomy (Human oversight)”: Such requirements also exist for non-AI-technology, i.e. for airbag control systems.5. Requirement #5 “Non-Discrimination”: Political impact of discrimination can lead to non-acceptance of AI in general.6. Requirement #6 “Respect for (& Enhancement of) Human Autonomy”: Possibility of situative opt-out is an important factor in this context (important for aware people, not so important for unsensible consumers).7. Requirement #7 “Respect for Privacy” should not be applied to AI-Systems specifically, but to the data used for input and training or to the data output by the systems. It should be stated that, with regard to privacy, AI-systems are just common IT-systems which should be subject to the privacy-by-design development paradigm as well as processes supporting privacy policies given by law. In our opinion the requirement can be eliminated from the list.8. Requirement #8a “Reliability & Reproducability”: Realisable for neural networks only if training sequence remains constant.9. Requirement #8b “Accuracy”: Depends on use-case.10. Requirement #8c “Accuracy”: Depends on use-case.11. Requirement #9 “Safety”: Only

						relevant for specific use-cases.12. Requirement #10 “Transparency”: Transparency could be a basis for official certification of AI-systems.	
Henning	Banthien	Plattform Industrie 4.0 (Germany)	General Remarks • The Plattform Industrie 4.0 welcomes the initiative by the HLEG to define ethical guidelines for trustworthy AI as an opportunity for Europe to provide a comprehensive framework for all human centered AI applications. • Often, fears and concern associated with AI are caused by the uncertainties around impacts of AI applications and the perceived potential loss of control. Thus, the Plattform is convinced that setting specific guidelines is a precondition for trusting and accepting the digital transformation, within and beyond industrial AI. • It is time to decide as a European community what kind of decision mechanisms we want to establish to prevent harmful scenarios of AI applications. Supporting this common effort, the Plattform Industrie 4.0 wants to emphasize the importance of striking the right balance between a precautionary approach and restrictions to the further development of AI technologies. The Plattform Industrie 4.0 calls for differentiated and detailed definitions and simultaneously considering AI technology and its industrial applications. The guidelines should be more specific on this. The industry perspective must be strengthened • The industrial sector plays a key role in the use and promotion of AI applications, with industrial AI contributing significantly to the GDPs of EU countries. AI is expected to contribute a third of the total forecasted growth in the German manufacturing sector over the next five years - equivalent to 32 billion euros. 25% of the manufacturing sector in Germany is already using AI (VDI/VDE 2018). It is our understanding that the industrial appliances of AI therefore should find their recognition in the Guidelines. • Europe has the opportunity to be an AI pioneer by building on the strength of the leading industrial sectors across European	Distinction between AI applications with a solely technological nature and those that involve human interactions • It is important to note that AI applications within the industrial setting are often technological procedures, focused e.g. on process improvement and flexibilization, optimization of cost, resource, time, energy or plant performance or automation of knowledge-based processes. It is important to dis-tinguish settings where AI systems have a significant impact on people’s lives and strictly technical processes. • Since industrial AI applications are predominantly of a technological nature without any human action involved the reference of an ethical purpose does not seem applicable und useful to be applied to these applications. This, however, does not mean these applications are not trustworthy in the sense of the guidelines. • Regarding “Traceability & Auditability” it is important to distinguish the con-text of application as well. When the systems do have a significant effect on people’s lives, laypersons should clearly be able to understand the causality of the algorithmic decision-making. In a technical process without effect on people life’s, experts can be the ones guiding and steering those applications. • Regarding the presented “Principle of Autonomy” it is noteworthy that decisions taken by machines typically follow task and goal-oriented algorithms. These algorithms and their goals have been programmed by humans. In contrast to deep learning of neural networks these can take place unsupervised. Efforts to define “Trustworthy AI” are welcomed. Thus, the Plattform Industrie 4.0 recommends to further clarify the premise in section B (A Framework for Trustworthy AI) that the definition of “Trustworthy AI” must aim at an ethical purpose. In this context, a "design for all" approach, as mentioned in the document, may not be adequate to ad-dress the broad variety of AI applications.	The development of Trustworthy AI must be feasible for all stakeholders, and duplication should be avoided • To mainstream Trustworthy AI applications widely across all sectors, bureau-cratic and technical hurdles as well as related costs should be kept at a min-imum. This is particularly important to SMEs that make up most of the sec-tor. SMEs typically do not have the capacities and resources to bear addi-tional costs, e.g. for assessment and monitoring schemes, which were mentioned in the guidelines. • To avoid unnecessary efforts and costs, it must be emphasized that ethical AI issues are already integrated within existing standardisation procedures, concerning e.g. smart manufacturing, robotics, autonomous transportation or cybersecurity. A duplication of these efforts should be avoided. The Plattform Industrie 4.0 is willing to transfer the general framework into domain and sector specific guidelines. The document should therefore be as specific as needed for a general setting. It should also clearly state, which measures are considered optional and which are compulsory.		The enabling environment of trustworthy AI applications should be addressed proactively • Given that the increasing use of AI technologies is profoundly changing the industrial sector, AI also needs to be integrated into qualification and lifelong learning schemes, to provide the necessary training and support for employees. The guidelines should stress this importance and give pursuant recommendations.

			nations. Especially with regard to machine learning and the use of other AI technologies Europe can play a key role, • To seize this opportunity, the guidelines should state the importance of creating a sustainable implementation framework for industrial AI application based on an accurate consideration of the technology. Consequently, it is our belief that the industrial appliances of AI therefore should find a stronger recognition in the guidelines.				
Patrick	Grant	BusinessEurope					As you know BusinessEurope delivered its comments yesterday on the AI HLG Guidelines. While we added many of constructive criticisms on elements where the text should be improved we realised that we failed to mention elements that are positive should stay in case others believe they should be rejected. Please find these points below and thank you again for your time and consideration. - The notion of AI having two components (ethical purpose/technical robustness) is a beneficial balance that indeed should encompass the entire guidelines - The idea that these Guidelines can influence and foster an ethical framework for AI at global level - Support for a tailored approach in AI (eg. different situations raise different challenges)
Anonymous	Anonymous	Anonymous	Being one of the most important Italian electronic communications companies, we welcome the opportunity to give feedback to the "Draft ethics guidelines for trustworthy AI". Our sector, in fact, is one of the most involved in the process of the development of AI, because in real life situations it might happen that sensors, actuators and AI agents are away from each other and one or more telecommunications networks lay in between and connect them. If this is the case, electronic communications network are clearly the infrastructure that transfers the information generated by the sensors to the remote AI agents. The new 5G and FTTH networks will, likely, be the main connective tissue of the artificial intelligence systems located throughout the territory.	We generally appreciate the intent of the draft guidelines to provide an overview of the fundamental rights and high-level principles and values with which AI should comply, set out concrete requirements (privacy, accountability and transparency) for AI systems. It is necessary to use policies on different levels to evaluate AI models. Policy developers including standards and regulatory bodies, should develop governance frameworks to make sure AI development does not infringe upon human rights, freedoms, dignity, and privacy. besides, starting with an overview of the issues and relevant ethical theories for AI, there will be the possibility of programming, using the AI solutions, co-robot ethics to legal and other questions, including liability and privacy concerns. This approach on trustworthy AI is aligned with our approach to offer solutions that people can trust, according to transparency and responsibility. This is also compliant to the approach followed by the AI HLEG [High level expert group on sustainable Finance -31.01.2018] also chose to start from fundamental rights, linking them to essential values, to define requirements for trustworthy AI. From our point of view, it is important to include in the final guidelines different use-cases, demonstrating how the guidelines can be applied in different AI contexts.	As a first feedback on the issue of guarantee of a trustworthy AI, it's important to consider the theme of the management of net neutrality when the use of electronic communication network is needed.In fact, the possible conflict in the network between special services and equally worthy of protection such as that given to an electro-medical device or a machine that can manage the braking system of a vehicle. Which of the two applications is more appropriate to give priority in case of problems? How is this decision made? To date, there is no answer and perhaps the only principle that can be given is that it should be established now, so that the rules are applied when we have millions of people who use these services.The first aspect to be defined is that of the relationships between the network operators, the software and hardware developers, deployers and the customer. Everyone has a central but different role, and everyone has their different responsibilities.The network indeed can experience congestion or outages and, when this occurs, the transferred information can be delayed or even discarded. Networks, in fact, are dimensioned using statistic criteria: it's impossible in fact to guarantee infinite network resources.- Two important safety properties that researchers have been studying and trying to give to the AI systems are known as "safe interruptibility" and "distributional shift". Safe interruptibility is the ability to interrupt an intelligent agent and override its actions at any time. On the other hand, distributional shift is the ability of the AI system to behave robustly when the environment where it operates differs from the training environment. - In our	Coherently to our feedback on chapter II, we believe that the assessment lists should be enriched by items to detect the correct application of "safe interruptibility" and "distributional shift", or other applicable procedures to avoid danger in case of outage of electronic communication networks.Further fine tuning should be applied on supervised machine learning algorithms and AI's essentially learn what they are taught. Therefore, for them to be fair and non-discriminatory, a proper choice of the training set is crucial to avoid imbalances and inequalities when classifying (i.e. with the term imbalance we mean that the classification results are not distributed evenly across all the classes forming the domain of the classification problem).Since miss-classification might lead to uneven decisions and bias reinforcement, an approach to mitigate the risk of miss-classification could consist in the assessment of the training sets in search of imbalances, not necessarily intended.Moreover, we generally agree on the principle that it's needed to guarantee not only the concept Privacy by design but also the concept of "ethics by design"; in this way we require ethical principles (and Privacy) to be embedded in AI products and services right at the beginning of the design process.Ethics must be embedded into the design and development process from the very beginning of AI creation and those ethics must be aligned with the values and ethical principles of a society or the community it affects. For this reason, it could be necessary to use the same approach in Europe to guarantee the process. Each company must share a common vision and	Artificial Intelligence requires an ongoing, interdisciplinary effort to cover all the effects in the different ecosystems where the new approach will be introduced.We think that the consequences of wrong decisions made by an AI are well diverse depending on the field they are made. Wrong decisions in a medical and transport context are clearly much worse than others. Therefore, we think that the regulation of the safety topic should be domain specific.Guidelines should give clear and unequivocal criteria to support an electronic communications network operator to make choices when network resources are insufficient to manage one or more mission critical services that transit networks simultaneously.In the same way, it is important to realise "Trustworthy AI" according to a shared framework. We look at the opportunities of enhancing data platforms in healthcare (through data that are anonymised and based on donor), or in other sectors so that AI can be trained to help improve diagnoses and treatments or other outcomes, like as smart mobility, smart cities or industry 4.0.

opinion, it must be sanctioned the principle that the network can't be held as responsible for any failure to achieve the two properties above (or similar). The reason why networks can't be considered responsible of the possible fallacies of the ensemble "sensors-actuators-AI agent" " is that zeroing the probability of congestion would require infinite network resources, which is clearly impossible . A fallback strategy for the AI agent in this scenario is in charge to developers, expected "by design". The AI equipment should sense the "hiccup information" of the surrounding congested networks and swiftly stop what it's doing, thus applying the principle of safe interruptibility.This applies even more when the consequences of wrong decisions made by an AI are well diverse depending on the field they are made. - Guidelines should give clear and unequivocal criteria to support an electronic communications network operator to make choices when network resources are insufficient to manage one or more mission critical services that transit networks simultaneously.In the event an inconvenience is unavoidable anyway, e.g. when humans are damaged due to a decision taken by AI systems, AI system must decide with fairness. Our second feedback on Requirements of Trustworthy AI concerns the correct assessment of the security measures to be provided for AI systems: it is necessary that they are previously subjected to a risk analysis based also on their intended use: the measures to be implemented must be adequate to the critical levels of the scenarios of use of the systems themselves (education, vehicles, healthcare, war,…) In particular, it might be useful to define a classification of AI systems according to their intended use in order to define the security measures adequate to the context of application. Could processes from other domains, such as whistleblowing protection and responsible disclosure, applicable for preventing AI-related misuse risks?Furthermore, in the requirement: "1. Accountability", an important role, in this context, is played by the promotion of awareness and culture of responsibility. AI and security experts are in a unique position in contributing to the correct use of the AI-enabled world, by providing awareness of threats and adoption of best practices.In the requirement: "8. Robustness" it would be appropriate to detail some concepts relating to security issues for the protection of the systems used to "train" the AI.It is also important to define security measures to ensure the availability and integrity of data throughout the life cycle of the entire AI system. Among the measures that help to ensure the robustness of the AI system is necessary to foresee the segregation of all environments including those provided for the AI training.In relation to sections "4. Governance of AI Autonomy (Human oversight)" and "8. Robustness", assuming the autonomy of an AI system as an axiom, the governance of this autonomy is fundamental to the security of the system and its scope of application (paragraph "Resilience to attack").Human supervision is only one of the possible instances of governance (see "Governance of Autonomy IA (Human Supervision)"). Taking into consideration the "different levels or governance instances" mentioned in the document, one could evaluate the

follow an internal procedure to guarantee Principles for Trust and Transparency. By adopting and practicing this approach it could be more clear and transparent how to use the AI solutions in different contexts Important points are as accountability, fairness, and explicability are considered in the document.Finally, consistent with our comments to Chapter II, we hope that in case the AI HLEG decides to introduce the concept of risk analysis as described in the comment to Chapter II, it would be appropriate to include questions about risk assessment in the assessment list.

				hypothesis of the aid of other AI systems to support human supervision, which carry out extemporaneous (therefore unpredictable) redundancy checks in environments separated by AI system to be governed, in a logic of "assisted governance". The autonomy of the AI system also depends on its ability to interact.The concept of governance then extends to the autonomous interaction of the AI system with other systems networks, devices and people close to its scope), on which contextually its evolution and its vulnerabilities depend. The system could be equipped with the capacity to autonomously expand its interactions within a predefined "evolutionary perimeter" during the design phase.The AI, equipped with its own start-up of a minimum set of connections to protect potential vulnerabilities, could independently activate further systemic interactions to increase its evolutionary degree. The improvement in efficiency would correspond to a reduction in risks related to vulnerability.	
Jessica	Cussins Newman	Future of Life Institute	I am writing on behalf of the Future of Life Institute (FLI), an international non-profit working on the benefits and challenges of emerging technologies including artificial intelligence. These Draft Ethics Guidelines are an important, concrete step forward in the international debate on AI ethics. In particular, the list of technical and non-technical methods and the assessment list will be useful to researchers and technology company employees who want to ensure that the AI systems they are developing and deploying are trustworthy. The report requests specific input on Section 5.5. We wish to share with you the longer-term concerns that have been voiced by thousands of prominent, international, AI and robotics researchers. In the course of your deliberations, you may have reviewed the 2015 Open Letter (Research Priorities for Robust and Beneficial Artificial Intelligence) and 2017 Asilomar AI Principles. These important documents were developed at our 2015 and 2017 Beneficial AI conferences, respectively. Thousands of AI and robotics researchers have voiced concerns that touch upon Section 5.5. This includes senior European researchers such as Erik Brynjolfsson, Yann LeCun, Francesca Rossi, Ramon Lopez de Mantaras, Pierre Marquis, Maria Chiara Carrozza, Bernhard Schölkopf, Marek Rosa, and Klaus-Dieter Althoff. The 2015 Open Letter, signed by more than 8,000 people including many prominent European AI scholars and EurAI fellows, recommends “expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial: our AI systems must do what we want them to do”. This includes both shorter-term and longer-term research. The 2017 Asilomar AI Principles, signed by more than 3,700 people, include on longer-term issues:“19) Capability Caution: There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.20) Importance: Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.21) Risks: Risks posed by AI systems, especially catastrophic or existential risks, must be	

subject to planning and mitigation efforts commensurate with their expected impact." Within the international debate amongst AI researchers regarding longer-term concerns, there is expert disagreement about whether and when AI systems will have various more advanced capabilities (also shown by surveys of leading AI researchers, see Grace et al, 2017.) However, there is agreement that AI significantly more powerful than today's could have profound effects on the global economy, society, and politics. There could also be great risks of accidents or misuse by malicious groups. Thus, a precautionary risk-management approach suggests we should plan for the future, monitor progress, and try to learn more. The current Section 5.5 does not quite capture the state of the international debate amongst AI researchers and developers. In particular, we think the following points could be emphasized more clearly:- Differentiating between transformative AI, AGI, consciousness, moral agency, and recursive self-improvement, as these are distinctly separate concepts about which experts have varying degrees of certainty across variable timescales. For example, transformative AI could refer to a sufficiently capable narrow AI, and is a highly likely occurrence given commercial incentives.- Acknowledging that the road to AGI may not happen by a predictable course of scientific advancement. AGI development could proceed on a smooth (but possibly rapid) progression to increasingly more capable systems, or by a single major new discovery in AI research. In a constructive spirit, we would therefore humbly suggest the following edited section: "All current AI is domain-specific and requires highly trained human scientists and engineers to precisely specify its targets. In the future, even sufficiently capable narrow AI is likely to become transformative as it is integrated throughout a greater number of industries. However, over the coming decades, there will be continued technological progress. There is no consensus in the AI research community as to the upper limits of the capabilities of future AI systems, or when certain milestones may be reached, though many experts believe greater generality in AI is likely. The development of AI systems that are highly and flexibly competent across a range of domains would represent a profound historical change, comparable in scale to the Industrial Revolution. Several critical long-term concerns can be identified: safety risks of accidents involving very capable AI, and security concerns of misuse by state or non-state groups. A risk-assessment approach, therefore, invites us to keep the rate of progress under consideration and invest resources into planning for and managing such changes, and reducing our uncertainty about these longer-term concerns. Other long-term ethical concerns could include the development of Artificial Consciousness and Artificial Moral Agents. It is also quite possible that artificial general intelligence (AGI) would never develop consciousness or require moral consideration. Such systems would nonetheless be hugely influential and have profound social, ethical, security, and political implications. Because advances in AI could come at an unpredictable time and/or rate, and have a significant impact, it is important to consider governance

mechanisms and ethical frameworks for the AI of today that can also smoothly and robustly scale to manage more capable systems.” We hope that this proposed rewrite will be useful, and are very happy to discuss any part of our response. In addition to the comments above related to section 5.5, FLI would also like to provide a more narrow comment related to section 5.4: Lethal Autonomous Weapon Systems (LAWS). We are in strong agreement with the vast majority of the section and commend the authors for a well-written and nuanced description of a complex issue in a small amount of space. As you may know, FLI organized a “Lethal Autonomous Weapons Pledge,” which has been signed by more than 240 organizations and 3100 individuals, including Google DeepMind, the European Association for Artificial Intelligence (EurAI), and Informatics Europe. Pledge signatories have committed to “neither participate in nor support the development, manufacture, trade, or use of lethal autonomous weapons" due to the threat of destabilization, violence, and oppression from automating lethal weaponry. With this effort in mind, FLI suggests removing the single sentence from section 5.4 that states, "Note that, on the other hand, in an armed conflict LAWS can reduce collateral damage, e.g. saving selectively children." We believe this claim is both too ambiguous (reduce compared to what?) and too speculative for inclusion in these guidelines. To responsibly make such a claim, the guidelines would require a significantly longer explanation of LAWS, including necessary caveats about the hypothetical technological design of future LAWS, alternative weapons systems, the potential political context surrounding their use, and whether LAWS might decrease psychological inhibitions about using deadly force. Given the high-level focus of these guidelines, we recommend simply removing the sentence. We sincerely thank you for your consideration. Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2017). When will AI exceed human performance? Evidence from AI experts. arXiv preprint arXiv:1705.08807.

Sana	Ali	ITI	The Information Technology Industry Council (ITI) is the global voice of the tech sector. ITI members are global companies with complex supply chains around the world, we understand the importance of consumers being able to use technology seamlessly across borders every day. As both producers and users of privacy protecting and enhancing products, technologies and services, ITI hopes to see the most effective approaches applied to promote the beneficial development and use of artificial intelligence (AI) globally. Glossary and Executive SummaryWe appreciate that the HLEG notes the incomplete nature of the glossary (pg. iv) and expects to further complement it. Below, we lay out our views about some of the current definitions.Artificial Intelligence (AI)We find that the definition for AI (pg. iv) is not in line with that generally used by the community of AI practitioners and the current state of art in AI technology. In particular, the statement that AI systems are "designed by humans" and are "deciding the best actions to take (according to pre-defined parameters) to achieve a given goal" appears outdated and ignores the existence of machine learning systems that are in fact not completely pre-defined by humans.There is a discrepancy in the definition of 'trustworthy AI' in the glossary section and the one used in the executive summary. We find the one in the glossary section is superior in that it makes clear that fundamental rights and regulations should be complied with during the development, deployment and use of AI (it is not the AI system itself that does all these things). We also endorse the acknowledgment that no legal vacuum currently exists, as Europe already has regulation in place that applies to AI. We believe building trust also means demystifying some of the unfounded concerns around the technology and educating the public on what AI is and how it can be used, and we recommend including these points. BiasThe definition of bias (pg. iv) is not aligned with its actual scientific meaning in statistics, instead overly focusing on the human element. It also overstates the risks compared to the advantages of AI, only once mentioning the potential for AI systems to support less biased decisions. We advise the following changes:"Bias is a prejudice for or against something or somebody, that may result in unfair decisions. It is known that humans are biased in their decision making and that unfair bias permeates our societies. Since AI systems are designed by humans and rely on data, it is possible that their results are biased even in an unintended way. Many current AI systems are based on machine learning data-driven techniques. Therefore, bias can manifest itself in the collection and selection of training data. If the training data is not inclusive and balanced enough, the system could learn to make unfair decisions. At the same time, AI can help humans to identify their biases, and assist them in making less-biased decisions."Endorsement MechanismWe commend the HLEG for acknowledging that a domain-specific ethics code – however consistent, developed, and fine-grained future versions of it may be – can never function as a substitute for ethical reasoning itself, which must always remain sensitive to contextual and implementational factors, and that different situations raise different challenges. Given the Guidelines'	Voluntary Rights Based ApproachWe are concerned that throughout this chapter, the voluntary nature of the Guidelines is not properly reflected. In particular it speaks of "governing the "ethical purpose" (pg. 5) and "identifies the requirements for trustworthy AI" (pg. 8) rather than providing guidance. There are also various instances where it is insinuated that the technology sector has a negative attitude towards their customers or individuals. The paragraph on respect for human dignity, for example, suggests that it be a requirement that "people be treated with respect due to them as individuals, rather than merely as data subjects"(pg. 9). This is not a fair representation and ignores that AI is actually used and developed by our companies to better fulfil individual needs. Finally, the language used in this chapter often overlooks that AI systems do not only "hold the potential to improve the scale and efficiency of government […] services" but they "are already improving" them.Implementing MeasuresWhen discussing the Principle of Autonomy to "Preserve Human Agency" (pg. 9), a right to opt out and a right of withdrawal are contemplated, but not qualified according to the use case. These rights need to be qualified for instances where opting out might cause harm to others or prevent an authority from performing its duties for the common good. such a 'right' can't be horizontal - it must vary according to the use cases.Similarly, when contemplating citizens' "right to be informed of any automated treatment of their data by government bodies" and "systematically be offered to express opt out" (pg. 7) the Guidelines overlook that most, if not all, government service provision will in future entail some degree of automatic processing of data and it is unclear how citizens would be informed of each of these. Secondly, and of greater concern, is that it is entirely unclear how an opt-out would work in practice. Would citizens, for example, have the right to have their tax declarations or social benefits allocations checked manually? This would undermine any smart governance systems and may result in uneven (and unfair) distribution or provision of public goods and services, not only for the citizens who opted out but also for those who did not.Furthermore, the Guidelines discuss rights associated with "direct or indirect" AI decision making- "a right to knowledge of direct or indirect interaction with AI systems, a right to opt out and a right of withdrawal". How this would work in the real world, even with today's use of technology, is entirely unclear. Ultimately, the only choice for the individual may be to not use the service at all. A similar lack of clarity exists around what is envisioned as being "effective redress if harm occurs" (pg. 10), and around why this should be singled out from a general right to redress. In addition, while measured transparency is indeed a key element in creating trust in AI systems, the Guidelines require "AI systems [be] intelligible by human beings at varying levels of comprehension and expertise" (pg. 10). This should only be required in qualified cases depending on the application and based on agreed procedures. The Guidelines also require "both technological and business model transparency" (pg. 10). Not only is the concept of 'business model transparency' unprecedented, it is also unworkable since	The HLEG notes that an agreement had not been reached on operationalizing the principles laid out in the first section and therefore sought input from the consultation. While the high-level principles and values laid out in the previous chapter are uncontroversial in and of themselves, we are concerned that they are not sufficient and further developed thinking is required on realising the goals stated therein. Many of our members have devoted thought and resources to developing more comprehensive internal guidelines, built specifically for developing and deploying intelligent systems. AccountabilityThis chapter aims to offer guidance on implementing trustworthy AI, but the description of "accountability" (pg. 14) is extremely narrow. It excludes preventative and systemic accountability processes within organisations developing or deploying AI systems, i.e. measures to ensure things do not go wrong in the first place and that, if something goes wrong, there is a procedure to follow and people in charge to address issues. Accountability should also include providing to the ability to contest AI output and provide feedback on why a certain output is right/wrong.Data Governance The data governance section is silent on many established best practices in data governance and handling, and instead is largely focused on the quality of data sets. Given this, this section should be retitled "data quality and governance". It should also devote some attention to the traceability of data sources, how data undergoes transformation, and maintaining documentation on the quality and nature of data, including considerations of potential re-identification of individuals. This section assumes that biases can be "pruned away before engaging in training" (pg. 14) while this may not always be possible and contradicts a later assertion that "data always carries some kind of bias". Machine bias can be introduced at various stages, be it due to characteristics of the AI's connectivity, the system's technical architecture or design, or through training bias (which is impossible to eliminate completely). Thus, bias cannot be removed or prevented, but it can be assessed, documented, mitigated and/or disclosed. One key consideration in implementing trustworthy AI is making this information available in a meaningful manner and determining what level of bias is acceptable for which application. We, therefore, suggest reframing to say that datasets inevitably contain biases, and one has to prune these away to the maximum extent possible before engaging in training. We also caution that suggestions to "always keep record of the data that is fed to the AI systems" (pg. 15) may not be always compatible with EU data protection laws.Governance of AI AutonomyWe commend the HLEG for their balanced approach in recognizing that assuring properties such as safety, accuracy, adaptability, privacy, explicability, compliance with the rule of law and ethical conformity heavily depends on specific details of the AI system, its area of application, its level of impact on individuals, communities or society and its level of autonomy (pg. 15).Respect for (& Enhancement of) Human Autonomy The autonomy principle is discussed at length in the fundamental rights section and is a much	In absence of a generalized model, the suggested approach to developing a set of guidelines involves assessing one use case at a time. We are generally supportive of this contextual approach, though some of our members have advanced a generalized model for assessing AI systems (independent of their nature) organized around types of biases. They have identified three types of biases that are detectable during assessment of any AI system that will lead to untrustworthy AI system operations: 1) biases related to data used to build the intelligence of an AI system, 2) human biases injected into knowledge bases gathered and used by an AI system, and finally 3) biases related to an AI system learning from another AI system. The HLEG might consider this level of abstraction in assessing AI systems.	General CommentsWe thank the European Commission High level Expert Group (HLEG) for their work on these guidelines and for taking a constructive approach that focuses on practical suggestions that AI developers and users can benefit from. We are happy to see that the Commission and the HLEG sees AI as a net positive for society and support the view that AI cannot be approached with one-size-fits all rules and prescriptions - but requires a constant discussion and iteration. We hope the HLEG will continue working this understanding into future steps of the work plan. Our general recommendation regarding the guidelines is that elements of it would benefit from further collaborative development, so they can be implemented and offer real world value. We have pointed out these areas and offered suggestions in the comments that follow.In addition, as AI is not developed in regional siloes, we recommend avoiding references to 'AI made in Europe', which are in contradiction with the global perspective the European Commission has endorsed. Products and services are the combination of components developed in different locations and are part of a global ecosystem. Many AI tools can be accessed via cloud computing and will work in combination with European and non-European elements. The Commission and the HLEG should aim to promote the ethical development & use of AI globally via collaborative engagement with its international partners. To realise ethical AI, the HLEG should consider the inclusion of a section on global governance. AI and technology as a whole are often built and applied across borders. They are part of an ecosystem, where different components might stem from different regions in the world. For this reason, the EU must maintain a dialogue with other geographies when it comes to the responsible development of AI. The most reliable way for Europe to ensure trustworthy AI for its citizens is to collaborate to promote a shared understanding and common norms across geographies. Europe should not miss the opportunity to shape the global debate on AI governance.

voluntary nature, we urge the HLEG to clarify the meaning and implications of the formal stakeholder endorsement process (pg. 2), as the current understanding may suggest some form of legal compliance and raises the question whether the guidelines may subsequently be referenced elsewhere (e.g. endorsement as a requirement in procurement procedures).

business models change over time and a system developed for one purpose could end up being used for another. These Guidelines should also make explicit that they similarly do not aim to imply disclosure of source code or any other information that would threaten industrial property or trade secrets.The Guidelines also refer to 'informed consent' (pg. 11), with reference to the GDPR though its meaning in this context is not clarified. We suggest the definition be specified, but with the additional consideration that GDPR allows data processing based on reasons other than consent, like legitimate interest. In many instances, a blanket right to refuse being subject to AI technology, would neither be possible nor desirable as it could go against the benefit of the user, against the rights of others or impede the functioning of public institutions. We suggest removing this concept altogether and replacing with the focus of these guidelines: trust. Finally, when contemplating Ethical AI, purpose and context go hand in hand. One must also be mindful of what practices are harmful and not harmful, lawful and unlawful. For example, the section on Identification and Consent (pg. 11) does not consider that not all identification processes create a danger for the individual and many are actually beneficial.

more expansive concept than what the draft has aimed to operationalize here. As discussed in Chapter I, human autonomy comes into play in many more contexts than B2C personalization online. Personalization is not only more complex, but also this kind of personalization could in fact augment human autonomy, rather than compromise it. In the right kinds of applications, AI could enable people to exercise much more precise preferences than would be otherwise practically feasible. We suggest the HLEG refer to the autonomy principle as discussed in Chapter 1 and simplify this section to say, "systems that are tasked to help the user, must respect their right to human determination, ensuring that the overall wellbeing of the user as explicitly defined by the user her/himself is central to system functionality" (pg. 17).Respect for Privacy The section on "respect of privacy" appears underdeveloped given the importance of this subject, and portrays data controllers as nefarious actors looking to "take advantage" (pg. 17). This is unwarranted, since as the HLEG has acknowledged, pre-existing regulations including the GDPR provide robust protections in this area.RobustnessThe section on "robustness and accuracy" (pg. 17) should further emphasize maintaining transparency about the level of confidence with which predictions are made or the level of uncertainty involved in those predictions.TransparencyThe "transparency" section overly focuses on the perception that one has to look into the "black box" (pg. 18). Given that this type of transparency may not always be possible due to the complexity of systems or their nature (e.g. self-learning systems), it is important to focus on the input and, even more, on the output stage to foster transparency. We also suggest the draft modulate the requirement of providing information about decisions concerning data sources, development processes, and stakeholders depending on the impact of the model on human beings. We also note that the term "human data" (pg. 18) is unclear and leaves little room for the nuance encouraged elsewhere in the guidelines.Traceability and AuditabilityThe "traceability & auditability" section (pg. 20) provides no parameters for what "transparent" and "understandable" could mean, which is key to having practicable guidance.StandardizationThe "standardization" (pg. 21) section appears to overly focus on standardizing the design of AI systems, rather than APIs and interfaces. It is unclear what the desired goal is for this kind of standardization, especially given that the nature of AI makes it difficult to imagine a horizontal standard that would be meaningful across applications and sectors.Codes of Conduct We suggest broadening this section (pg. 22) to include other modes of self-regulation.

Anonymous	Anonymous	Anonymous	“A mechanism will be put in place that enables all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis. This will be set out in the final version of the document.”In order to ensure trust it must be ensured: In any case, if a company publicly states that it adheres to the guidelines, in order to attract consumer trust, but in practice, it doesn’t respect its principles, this should be considered as an illegal be-haviour and sactioned accordingly. A general “absolution” from legal consequences is likely to undermine consumers’ trust in the Ethics guidelines in general.	4. Ethical Principles in the Context of AI and Correlating Values The Principle of Autonomy: “Preserve Human Agency” addresses responsibility and accountability: “It is paramount that AI does not undermine the necessity for human responsibility to ensure the protection of fundamental rights” At the same time LEGAL accountability i.e. liability is not explicitly addressed in the ethical val-ues. 5. Critical concerns raised by AIThe chapter on “Critical concerns raised by AI” addresses important aspects of AI. The guidelines should address the risks and benefits of AI-related technology. Omitting or diminishing these critical aspects in the final guidelines poses the risk of introducing a bias towards a naively optimistic description of the of AI-related dis-cussion. The same holds for the paragraphs on the “Potential longer-term con-cerns”	It should be incorporated that an independent external control system/Institution should be able to access (e.g. via APIs), review and audit AI/ADM processes. 1. Requirements of Trustworthy AI1. Accountability The necessity of humans being LEGALLY accountable i.e. liable should be ad-dressed explicitly.2. Data GovernanceIt should be addressed, that an independent external control system/Institution should be able to access (e.g. via APIs), review and audit AI/ADM processes, e.g. with respect to bias.5. Non-DiscriminationIt should be addressed, that an independent external control system/Institution should be able to access (e.g. via APIs), review and audit AI/ADM processes, e.g. with respect to compliance with anti-discrimination laws. This can be achieved by so-called input-output tests when an API to the database and the outputs of the AI Systems are provided. 8. Robustness as well as Testing & ValidatingIt should be addressed, that an independent external control system/Institution should be able to access (e.g. via APIs), review and audit AI/ADM processes, e.g. with respect to the robustness sof the Database, methodology and models em-ployed in a AI system. Documentation of the conducted tests should be made ac-cessible to the audit team for external validation.StandardisationStandards must be developed in order to allow an independent external control sys-tem/Institution to access (e.g. via APIs), review and audit AI/ADM processes. That entails also proper documentation of the AI-processes that enables the audit team effectively auditing the system.	An independent external control system/Institution should be able to access (e.g. via APIs), review and audit AI/ADM processes. This should be reflected in the as-sessment list.1. Accountability The necessity of humans being LEGALLY accountable i.e. liable should be ad-dressed explicitly.2. Data GovernanceIt should be addressed, that an independent external control system/Institution should be able to access (e.g. via APIs), review and audit AI/ADM processes, e.g. with respect to bias.5. Non-DiscriminationIt should be addressed, that an independent external control system/Institution should be able to access (e.g. via APIs), review and audit AI/ADM processes, e.g. with respect to compliance with anti-discrimination laws. This can be achieved by so-called input-output tests when an API to the database and the outputs of the AI Systems are provided. 8. Robustness as well as Testing & ValidatingIt should be addressed, that an independent external control system/Institution should be able to access (e.g. via APIs), review and audit AI/ADM processes, e.g. with respect to the robustness sof the Database, methodology and models em-ployed in a AI system. Documentation of the conducted tests should be made ac-cessible to the audit team for external validation.	Independent and external Control of AI needed to create trustThe Draft guidelines are unfortunately do not adequately address one aspect that is vital in making an AI system trustworthy: The establishment of an independent con-trol system that is able to access and audit AI/ Algorithmic Decision Making Pro-cesses (ADM processes). The adoption of principles with respect to the design and employment of AI by organisation/companies cannot create trust among consum-ers itself, if adherence to these principles cannot be independently verified/falsified by an independent external audit. Two important factors can contribute to create ´trust and acceptance could be promoted A) when automated decisions become transparent and explainable (which is ad-dressed in the guidelines) b) when there is a proper control system in place, that makes sure that the deci-sions about consumers are lawful, ethically sound and based on a rigorous meth-odology/database.The establishment and enablement of an external control/Audit system is not suffi-ciently addressed in the guidelines. It should be given a prominent place in Chapter II: Realising Trustworthy AI as well as Chapter III: Assessing Trustworthy AI.Such an external audit should test whether the system conforms with legal re-quirements: anti-discrimination law, unfair competition, data protection and it should analyse individual and social impact of AI. In order to establish trust, we need to ensure that these black-box systems are designed in a way that allows them to be independently accessed, controlled and audited, so that they comply with legal requirements and can be evaluat-ed/audited by experts, e.g. with respect to their social consequences. (e.g. with respect to consumer protection laws, anti-discrimination and data protection laws) In order to facilitate an audit in the first place we should consider establishing standards for transparency-by-design and accountability-by-design. These standards could ensure that third party experts get access to meaningful infor-mation (e.g. standards for documentation, APIs to test whether the database or outputs are biased). The focus of such tests should be on socially relevant ADM/AI processes. S those that potentially affect many consumers or have large adverse effects on them.

Anonymous	Anonymous	Anonymous	Facebook appreciates the opportunity to provide comments on the draft Ethics Guidelines (hereinafter, "Guidelines") of the High-Level Expert Group on Artificial Intelligence ("AI HLEG"). AI powers much of what we do at Facebook. On a daily basis, Facebook works to advance the field of machine intelligence and to create new technologies to give people better ways to communicate. Our research arm, Facebook AI Research ("FAIR") seeks to understand and develop systems with human-level intelligence by advancing the longer-term academic problems surrounding AI. FAIR covers the full spectrum of topics related to AI, and to deriving knowledge from data: theory, algorithms, applications, software infrastructure and hardware infrastructure. Applied Machine Learning ("AML") is essential to Facebook and connects our efforts between research and experiences on our platform. Furthermore, Facebook supports independent and cutting edge research on fundamental issues impacting AI, such as safety, privacy, fairness, and transparency. Facebook has partnered with the Technical University of Munich (TUM) to support the creation of an independent AI Ethics Research Institute. Drawing on expertise across academia and industry, this Institute will conduct independent, evidence-based research to provide insight and guidance for society, industry, legislators and decision-makers across the private and public sectors. Furthermore, Facebook recently collaborated with the Digital Ethics Lab of the University of Oxford to assess, map and explore how AI can help meet the United Nations Sustainable Development Goals. In sum, Facebook seeks to harness the power of AI to promote human understanding and wellbeingTo that end, Facebook supports the goals of the HLEG to ensure that AI is developed and used in a responsible and ethical manner. It is critical that we build AI systems that are safe, work as intended and are guarded against adversarial manipulation. At Facebook, we are committed to the development of responsible and trustworthy AI, using an open model that encourages collaboration on these goals. We also know that we are at the very early stages of AI technology, and that the global community working on AI should strive to promote innovation that makes progress towards solving our greatest challenges. Below we offer more detailed views on these perspectives as well as suggestions towards improving the Guidelines.	Facebook is committed to using and developing AI in a responsible and ethical manner. Facebook works with leading industry voices, academics, and policymakers as part of the Partnership on AI to Benefit People and Society, which studies and formulates best practices on AI technologies. The Partnership on AI has put forth central tenets which aim to promote fairness and inclusivity, explanation and transparency, security and privacy, values and ethics, and trustworthiness, reliability, containment, safety, and robustness of the technology. Additionally, we are part of the AI4People Initiative, the first global forum in Europe on the social impacts of artificial intelligence. This consortium, which is comprised of representatives of governments, European institutions, civil society organisations, and leading businesses, is tasked with designing a European ethical framework for a "good AI society".To that end, Facebook supports the goals of the Guidelines to create an ethical framework to achieve "Trustworthy AI." We strongly agree that AI must be developed, deployed and used in a human-centric manner, governed by an ethical framework that reflects fundamental rights and societal values. At Facebook, we take a holistic view in what comes to AI and Ethics: investing in the people who are building the algorithms, the data we use to teach AI, and the algorithms that represent what the AI ultimately learns — all feeding into the core technologies we design and deploy. Because it's people who design, develop and generate the data that teaches AI, we need to understand and mitigate our biases to ensure we don't pass them to the AI we create. To help do this at Facebook we have research, product and other review processes that act as independent checkpoints on the work people are doing. These involve external feedback which helps grow the table and ensure that a multitude of inputs shape our direction. We also consulted with leading researchers from the algorithmic fairness community and developed a new internal tool called Fairness Flow. Fairness Flow can help generate metrics for evaluating whether there are unintended biases in certain models. With this in mind, Facebook applauds the HLEG's inclusion of criteria for trustworthy AI that promote justice and fairness. We agree that AI should not only respect but develop further the fundamental rights of human beings and support the common good. To achieve these aims, we offer our suggestions below on the HLEG's proposed ethical framework for trustworthy AI.Introducing two additional dimensions to the definition and framing of "Ethical Purpose"Ethical purpose, as defined in the Guidelines, should not be subsumed to legal compliance, but strive to go beyond it. Ethical purpose should also recognize the role of the values and principles that actors developing AI systems have publicly disclosed as their own internal principles and mission statement, as the latter are key in addressing complex decisions and trade-offs that need to be made within product development processes. We thus encourage the HLEG to integrate these two important components in its current framing and definition of ethical purpose. Starting with the first, this has been defined in the academic literature as "post-compliance	An important element to achieve trustworthy AI is user control. It is important to make sure that users have control over automated decisions as much as possible, for instance they should be able to say that they don't want ads that entice them to drink alcohol, buy sugary drinks etc. Moreover, there is a need for a rigorous software development process with review. An additional section in the 'Non-technical methods' part could be added, calling for a rigorous risk-based development process with checklists, tooling and review.	The Assessment List seems slightly unstructured and repetitive. Given that the list will be substantially changed to match the four use cases that will be added to the next draft, we will be happy to add more specific comments in the next version. Our main suggestion at this stage is to tie the assessment list to a clear and operational risk assessment methodology, which is lacking in these Guidelines and that, in our opinion, would contribute to the operationalization of the principles and values listed in the document.	We appreciate the opportunity to provide our comments on the Guidelines. Facebook shares the goals of the HLEG to develop and implement a framework for AI that ensures that the technology is trustworthy and robust, and—most of all— advances our progress towards solving humanity's greatest challenges. We encourage the HLEG to continue to seek industry engagement in this discussion of how to promote responsible development and use of AI and look forward to continuing to work together towards our shared aims.

ethics”: ethics that elicits what can be done over and above what legislation strictly requires. This is not ethics against law, or despite its scope, or designed to change or by-pass it (Luciano Floridi “Soft Ethics and the Governance of the Digital”). The ethics as beyond compliance was also identified as one of its core features at the 40th edition of the International Conference for Data Protection and Privacy Commissioners (ICDPCC), devoted to the theme “Debating Digital Ethics”. As stated in the Conference's website (https://www.privacyconference2018.org/en/conference/ethics): “ Beyond compliance does not mean beyond the law or as an alternative to the law. The law must of course be complied with. But as we are seeing, compliance alone cannot preserve our rights and values.”When building products that leverage AI technology, there will often be tensions and conflicts between values and principles. Ethics is not only about following principles; it's about balancing them, attaining the right trade-offs and compromises. This means that digital ethics must serve a practical function of helping practitioners make decisions when operational choices implicate striking an appropriate balancing between competing values. As a framework for decisionmaking, it must enable AI practitioners to weigh and apply internal or organizational values and principles to effectively inform final decisions. AI developers and deployers should also leverage digital ethics when there is more that can be done over and above what legislation strictly requires. For the latter, and in the case of Facebook, this includes projects on computer vision for the blind community, partnerships with humanitarian organizations to help respond to natural disasters, AI open source projects, academic research and initiatives around open communication and transparency about difficult decisions we are called to make (“Hard Questions” blogpost series), and development of AI educational videos, amongst others.We thus recommend the HLEG to incorporate in its framing of the ethical purpose the post-compliance nature of ethics, alongside with a reference to AI stakeholders' own internal values and principles.The Guidelines allude to the tensions that may arise between principles: “It should also be noted that, in particular situations, tensions may arise between the principles when considered from the point of view of an individual compared with the point of view of society, and vice versa.” (p.8) In those circumstances, the HLEG recommends revisiting the principles and overarching values and rights protected by the EU Treaties and Charter, along with resorting to internal and external ethical experts. While we do agree with these recommendations, we suggest adding to them an acknowledgement of the post-compliance nature of the ethical reasoning process, along with a reference to institutions' own principles as a way to address possible tensions between competing values and inform final decisions.Emphasizing the need to assess the scope and coverage of AI by existing regulatory frameworksWe support HLEG's emphasis on the need to acknowledge and leverage existing regulatory frameworks that already apply and cover AI. Indeed, we should first apply and assess existing policy

and regulatory frameworks to new technologies, rather than trying to regulate them distinctly or from scratch. It's essential to assess the need for regulation within the full context of existing frameworks and with a balanced perspective of the specific concerns raised by algorithms with the many current and future benefits they offer to society and our economy. The Guidelines allude to the importance of leveraging current regulation: “it should be noted that no legal vacuum currently exists, as Europe already has regulation in place that applies to AI ...”. We urge the HLEG to continue stressing this component in the overall AI Governance debate. II. Facebook encourages HLEG to develop guidelines that promote innovation and the potential of the technology to solve humanity's biggest challenges.We agree wholeheartedly with the HLEG in its definition of human-centric AI that human values should remain the primary consideration and that the development and use of AI should not be seen as a means in itself. Rather, the goal of AI should be to increase the well-being of humanity.This primary focus on the value of AI to humanity, rather than the technology itself, requires that we approach these guidelines in a careful manner. We sit at the beginning of the AI revolution, and we have yet to scratch the surface of its potential. Thus, any recommendations or standards governing the technology should, at this stage, should seek to foster innovation and progress. Overly burdensome or prescriptive guidelines— before the technology, use cases, or even its conceptual potential have progressed beyond this nascent point— fundamentally would conflict with the goal of human-centric AI to increase human wellbeing. In other words, we should be cautious not to propose guidelines that may forestall future breakthroughs. To that end, we offer our suggestions below.Elaborating on the proposed AI tailored approach and implementing a risk based approachWe fully endorse the HLEG's recommended tailored approach to AI by emphasizing its' context specificity. It is vital to acknowledge the different types of automated decisions that AI can facilitate, its different implications, sensitivities and impacts on people, and - in that way - avoid unnecessary generalizations that will only end up hindering or blocking the development of this technology and the fulfillment of its many benefits. We encourage the HLEG to elaborate upon this tailored approach and delineate the foundations for a taxonomy and benchmark that provides guidance and clarity to all actors involved in designing, developing and implementing AI systems.[p.3: “ AI systems recommending songs to citizens do not raise the same sensitivities as AI systems recommending a critical medical treatment. Likewise, different opportunities and challenges arise from AI systems used in the context of business-to-consumer, business-to-business or public-to-citizen relationships, or – more generally – in different sectors or use cases. It is, therefore, explicitly acknowledged that a tailored approach is needed given AI’s context-specificity.”] We believe the AI Guidelines would benefit from the endorsement of a risk-based approach that will look, ponder and assess the harms and the benefits of the AI systems being built and deployed. Such a risk-based approach would render the operationalization

of the ethical purpose more palpable and actionable. Along these lines, we also urge the HLEG to connect more explicitly its proposed AI tailored approach to some of the arguments presented throughout the Guidelines. For example:- when advocating for the need for AI to "provide due process by design, meaning a right to a human-centric appeal, review and/or scrutiny of decisions made by AI systems" (p.7),- when articulating a "right to decide to be subject to direct or indirect AI decision making, a right to knowledge of direct or indirect interaction with AI systems, a right to opt out and a right of withdrawal" (p.10)- when defending the need for AI developers and deployers to ensure "that humans are made aware of – or able to request and validate the fact that – they interact with an AI identity" (p.11)- when listing the need for a process to allow human control in each stage of the development process (p.24, within the assessment list under the "Governing AI autonomy" section).It is important to avoid broad generalizations and - truthful to the AI tailored approach - condition the application of those rights and requirements to the nature (sector, context, purpose) of the AI system at stake and to a specific harm assessment threshold and benefit factors, where only automated decisions that severely or significantly impact individuals would be subject to higher levels of scrutiny. The Guidelines do reference the need for harm to occur as a justification for users' effective redress when articulating the "Principle of Justice: "Be Fair". We encourage the HLEG to elaborate further on this idea, implementing a risk based approach into the Guidelines.And here we would reference the AI Model Governance Framework that the Singapore Personal Data Protection Commission has recently released for publication, which contains specific guidelines on probability and severity of harm, along with human in - out - and above the loop scheme as criteria for informing its risk-based assessment to AI governance. In addition to the above areas, Facebook encourages the HLEG to consider openness and collaboration as a key aspect for ethical development of AI technology. Here at Facebook, our teams publish research results and open-source our cutting-edge research code, data sets, and tools. FAIR has applied an open model to all aspects of its work, collaborating broadly with the international research community and publishing research results in various fora.This free movement of research and open exchange of ideas allow for global review of risk areas and opportunity. Working in the open allows everyone to make faster progress on AI, making it easier to have an informed debate around safeguards as more people and institutions make use of new technology and research. We believe that open work on new technologies improves the underlying science and the pace of innovation while encouraging the development of ethical and trustworthy AI. We encourage the HLEG to adopt an approach that supports this open model.Need for clarity regarding the propositions on Algorithmic Explicability and AuditabilityWe also encourage the HLEG to explain in more detail its position and recommendation on algorithmic auditability. The current framing is unclear towards who, when, under which criteria, for which types of automated decision making processes,

and under which risk/harm based assessment: “Technological transparency implies that AI systems be auditable, comprehensible and intelligible by human beings at varying levels of comprehension and expertise” (p.10). In the Assessment List (Section III of the Guidelines), one of the questions proposed that fall under the Accountability domain is about an (external) auditing of the AI system being foreseen? (p.24) We recommend looking into the feasibility of these potential external mechanisms and tying them to a risk based approach before integrating them into this specific assessment list.We also urge the HLEG to expand on the operational aspects of the Principle of Explicability and clarify how, in practice, would “Individuals and groups ... request evidence of the baseline parameters and instructions given as inputs for AI decision making (the discovery or prediction sought by an AI system or the factors involved in the discovery or prediction made) by the organisations and developers of an AI system, the technology implementers, or another party in the supply chain” (p.10). Given the complexity of some AI systems, it will be highly difficult (if not unfeasible) to provide such information. We also think that such technical and exhaustive approach may not be the most appropriate and meaningful to end users. We encourage the HLEG to clarify and simplify how it foresees the actual application of the principle of explicability. In addition, and in line with the prior observation, we insist on the need for a risk based approach in order to operationalize this explicabillity principle, applying it only to (consequential, harmful) automated decisions.Informed ConsentWe fully agree that informed consent is absolutely crucial as a value needed to operationalise the principle of autonomy, explicability and non-maleficence. Nonetheless, and given the inherent tensions emerging from the strict application of the collection limitation, purpose specification and use limitation principles in the AI context, we recommend flexibility in its interpretation and application. Given the largely automated learning process that informs and characterizes AI systems (namely Machine Learning), which may yield potentially unforeseen, but positive, and important, results, it is becoming increasingly difficult to continue operating on the basis of full knowledge and articulation of purposes. We thus recommend looking into ways to render the application of informed consent more attuned to these particular circumstances. We also believe that a risk based approach could help operationalizing the informed consent in more flexible way.

Joanna	Lopatowska	EuroCommerce	• When talking about the ethics of AI, ethic data collection should be mentioned and elaborated in the Draft. AI relies on data to learn and developers and users have the responsibility to select trustworthy data, which will eventually lead to trustworthy AI. • AI is just one of the many methods to evaluate data. The downstream automation of decisions based on analysis results is not AI specific. For all analytical procedures (including AI), the respective analysis result depends directly on the data provided. This means that if input data contains ethically reprehensible information, or if such information can be derived from the data, in principle these results can be made visible (independently of the method). AI significantly accelerates the finding of such results. However, AI per se, is not the cause of possible reprehensible results, but rather the data that are fed into the process. This also explains that with input data that are subject to a certain bias (for example, data that are predominantly involving negative assessments/aspects of a person), the analysis that will be derived from those data, will also follow the bias. The cause is not the analysis methodology, but the input data provided. The discussion should then also concern the underlying data and not just the analytical method. • It would be helpful if the Draft provided practical examples of how to achieve requirements of trustworthy AI. Further, it would be beneficial if the Guidelines would shed light on AI solutions that are already part of our everyday life – sometimes without people even noticing it – to break stereotypes and create grounds for trust.• Regarding the target audience, we would welcome clarification that the principles and spirit as enshrined in the Guidelines should drive the regulators – both at the EU, as well as at national level in case any measures are taken ahead of the EU action – when developing guidance, policies or measures concerning AI to secure a regulatory environment supportive of innovation that will safeguard society from the challenges Artificial Intelligence raises.• There needs to be more clarity between the terms "user" and "consumer" as these seem to be sometimes used interchangeably, which leads to confusion. We would suggest a consistent use of the term individuals, as opposed to business use of AI.	• Relying on the fundamental rights commitment of the EU Treaties and the Charter of Fundamental Rights and all associated stepping stones to best identify ethical principles is a powerful encouraging concept. • We would welcome guidance on how to address situations where AI solutions face conflicting rights in the Charter, as there is no hierarchy of rights in the Charter. • In particular, the Charter also includes the freedom to conduct a business, which should be taken into consideration when discussing the potentials of AI• The Guidelines would benefit from a strong emphasis that AI development must be conducted with the EU fundamental rights at the core and with these rights and values as a default option. For example, AI may never be used for discrimination, harassment and personal privacy must always be respected and AI must never be used so that it becomes conflicts of interest. • Respect for human dignity is one of the main fundamental principles that should always be taken into consideration when using AI. • In addition, the General Data Protection Regulation sets foundations for the legal and fair processing of personal data and this Regulation should also be an anchor for framing the Ethics Guidelines. Introducing new regulations covering personal data aspects related to AI, would create a risk that they are not consistent. • Regarding critical concerns raised by AI, more reflection is needed on the identification of individuals without consent. Article 6 of the General Data Protection Regulation requires valid legal basis as a precondition for data processing, and Article 9 sets specific grounds for the processing of sensitive data, including biometric data. However, consent is not the only legal basis laid down by Article 9 and identification is not only taking place via the processing of biometric data. We believe more reflection is needed on using AI for identification and the relevant legal grounds.	• We welcome the identification of clear requirements to achieve trustworthy AI. Nonetheless, examples of technical solutions to meet such requirements would improve the usability of the Guidelines for businesses. • In our view, transparency and accountability at each development step are the central concepts that should guide the implementation of trustworthy Artificial Intelligence. Transparency towards individuals (e.g., clearly mentioning when AI solutions are used along the sales chain) will increase the confidence in AI and the normalisation of the technology. • Too many requirements for trustworthy AI put at risk the fast development of European AI technologies. The European Union needs to strengthen its position vis-à-vis China and the United States. We would like enforceable requirements that apply to AI solutions regardless of where they developed. • Attention must be paid to the type and quality of data that is input in AI solutions to avoid discrimination.	• The introduction of an assessment list is very good and helpful for most of the actors on the market. We would suggest the following to strengthen this approach: • Regarding accountability, the Guidelines should acknowledge that, in practice, more than one actor is often responsible for AI across its lifecycle. The Guidelines should clarify who should be responsible for AI solutions across their lifecycle – the producer, the software developer, the user or all. • Regarding data governance, resources capacities to appoint a dedicated AI responsible should be taken into account given that not all companies will have the capacity to manage and finance such requirement. A more nuanced approach could stress that all the points in the assessment list can should be implemented in accordance with the prerequisites and capacities of each organisation. • Regarding design for all, a clarification on the meaning of "wide range" would be useful. • Regarding non-discrimination, the Guidelines should clarify how the principle of non-discrimination applies in the context of segmentation of individuals. Segmentation is not per se discriminatory. Businesses segment customers already today with AI to better understand customer preferences and better tailor offers. The Guidelines could set conditions under which customer segmentation in compliance with its ethical principles. • Regarding transparency, information to individuals should be meaningful as opposed to providing technical details that could create confusion. In addition, it might not be possible for a company to describe the technology behind the system since the business user of an AI solution is not always who developed it. • We would welcome a reference to digital education to be introduced into the final Guidelines of the High Level Expert Group to address the continuous need to invest in digital literacy. Digital skills are paramount to develop AI solutions and users need basic digital skills to understand the functions and impacts of AI. • Clarification is missing on the application of the Guidelines to the existing and used AI. For example, what should happen to AI programs that are already developed? Should they be reprogrammed or redesigned to fit the requirements of trustworthy AI? If this is the case, it would entail significant costs and delay the entire development process of AI.	EuroCommerce is the main European organisation representing the retail and the wholesale sector. It embraces national associations in 31 countries and 5.4 million companies, both leading multinational retailers such as Carrefour, IKEA, Metro and Tesco and many small family operations. We welcome the draft Ethics Guidelines for Trustworthy AI and the opportunity to provide our comments and have a voice in the debate. We welcome the document as a first step to launch an EU-wide debate on ethical issues around Artificial Intelligence (AI). We support the approach to start from the principles to (i) lay common foundations for the understanding AI's impact on individuals, corporations and competitiveness, (ii) develop common understanding of the challenges, and (iii) identify key guidelines. As companies are exploring how to integrate AI solutions in their business models, we believe it is still too early for engaging in regulation and we welcome that the Guidelines resist such temptation and pursue a cautious approach. Equally, we believe that in order to avoid confusion there should be caution in having too many soft law measures such as guidance. The retail sector is taking up AI successfully and is becoming one of the sectors that is fastest investing in such technologies. For example: • Two in five retailers and brands are already using intelligent automation, and adoption is expected to double by 2021;• Intelligent automation is making the biggest impact in demand forecasting, supply chain planning, customer engagement, transportation and logistics and in-store services;• Companies invest in intelligent automation with an eye towards improving efficiency and customer experience. However, as these capabilities mature, they realize additional benefits beyond their original expectations.Therefore, we commend the Guidelines for stressing that any possible measures should aim at creating the most favourable climate to AI innovation that will benefit people and businesses. We share the perspective and the recognition of AI's great potential to improve human lives and that it must be based on the respect for the fundamental rights. It is crucial for the EU to lead global ethical standards on AI.
Mathana	Stender		The definition of 'Bias' is inadequate: it- 1. focuses on humans and training data, and does not address parts of the algorithmic ecosystem like the choice of regression models, the human role in interpretation of results 2. does not create a scalable framework. A theoretical approach mapping the distance between variables and constants throughout the algorithm's 'life cycle' would better encapsulate the 'bias transfer' as data makes its way from input to output.	5.1 Consent: Add: M&As + National 'Critical Social Infrastructure' data holdings 5.1.1 A benchmark or framework to assess the potential impact of companies merging with- or acquiring other companies with vast troves of personal data holdings. Though there are currently regulatory frameworks to ensure fiduciary responsibility, ethical AI guidelines should require the auditing of databases to ensure that bias is reduced within the acceptable boundaries and ID-sans-consent arising from metadata acquisition matched against current data holdings (eg de-anonymization) is achievable. 5.2 Covert AI: This section should make a distinction between surveillance systems that pilfer data (ie CCTV cameras) and 'AI agents' with active agency (ie chatbots that try to 'up-sell' one into consumption	2.1: Traceability & Auditability: This section should more clearly delineate what sort of 'traces' an algorithm should be required to produce. I suggest the inclusion of the terminology 'artifact' as something that are generated at each 'step' of the way. This framing is also important for *forensic algorithmic analysis* 2.2 Standards: the IEEE is in the process of producing a number of (voluntary) standards around AI and autonomous systems. -ISO, 'Fair Trade' and Made in Europe standards models all have different verification, validation and enforcement mechanisms. It is important that a 'standards' section clearly lay out both *what* and *how* technical and non-technical standards are defined, assigned,		I give the authors the right to publish my comments with or without attribution of my name. I agree both that -my comments on the draft guidelines will be openly published on an anonymous basis (i.e. without mentioning my name and, if applicable, organisation). -my comments on the draft guidelines will be openly published with identification of my name and organisation.

			decisions) 5.4 LAWS: A blanket ban on technology transfer (material, technological or otherwise) from any EU member states to any external nation that develops LAWS. 5.4.1. Export controls: Increase the burden of proof on countries and other entities seeking to purchase AI technologies that could serve in a dual use (eg lethal and non-lethal) capacity - except in cases of continual, transparent auditing. As a drone navigation system could be used to provide autonomous navigation for a UAV with a lethal payload, such technologies (even if commercially available outside of the EU) should require additional oversight.	complied with and enforced.	
Anne	Joseph	RELX Group	RELX Group is a global provider of information and analytics for professional and business customers across industries. Our brands include Elsevier, LexisNexis, Accuity, FlightGlobal, ICIS, ProAgrica, MIPCOM and World Travel Market. By combining content, data, and advanced analytics we help doctors save lives, researchers make new discoveries, insurance companies offer lower prices and lawyers win cases. We save taxpayers and consumers money by preventing fraud and money laundering and help executives forge commercial relationships with their clients. In short, we enable our customers to make better decisions get better results and be more productive. Our open source big data technology known as HPCC (High performance Computing Cluster) is used to analyse structured and unstructured data giving our customers the information and insight they need. HPCC powers the world's largest public records database processing 30m transactions per hour, with over 8000 technologists employed across the group. Our analytics tools enable customers to manage risks, develop market intelligence, improve economic outcomes, and enhance operational efficiency. We support drawing up of ethical guidelines and welcome the Commission's initiative in forming the High-Level Expert group on Artificial Intelligence (AI). We look forward to supporting its work directly through the contributions of Dr Elizabeth Ling, recently appointed as a member of the group representing RELX, and indirectly through being members of the AI Alliance. We support the approach of focussing on two overarching strands of ethical purpose and technical robustness. We also strongly agree with the objective of the Guidelines to form a proportionate and risk-based approach to AI. We fully agree with the need to recognise that "one size does not fit all" and that specific applications will each have their unique requirements. Applications will be highly context specific and the Guidelines should more serve as elements to consider for appropriateness rather than as requirements to be implemented in all contexts. At the same time, there is a concern that a formal endorsement mechanism would potentially prevent these Guidelines from evolving. Build on existing principles	Principles need to be holistic The Guidelines refer to existing treaties and charters as stepping stones towards identifying abstract ethical values and principles. These concepts are well established and form a useful baseline. They fit well with the vision and mission of RELX products to deliver improved outcomes for our users across multiple sectors. As with any other technology tool, usage of AI should take account of the impact on different stakeholders. The Guidelines recommend the principle of "do no harm ". Whilst this principle should be adhered to in the vast majority of cases, there will be some instances, such as identifying fraudsters or sex offenders, where there is clear wider social benefit in taking a broader view: individual bad actors may and indeed should suffer negative outcomes. Also, ethical principles can conflict with one another and trading off between benefits and harms may be unavoidable. The notion of proportionality is key. For example, scoring is already well established in credit rating and insurance premium evaluation. This benefits individuals in giving them the best opportunity to access products at the fairest rate. The proposed "right" to opt out and withdraw from AI decision making processes in certain selected fields of application is not realistic and will be difficult if not impossible to implement. This could cause harm to others or prevent an authority from performing its duties for the common good.	Explainability Rather than setting rules for every aspect of AI in detail, there is a clear case for different levels of explainability for different types of AI. For example, in relation to imaging-based diagnostic AI, currently the most common use of AI in health, explainability may prove difficult to achieve. It may be difficult to trace back evidence points accurately in the identification of disease-indicating anomalies in an X-ray. In circumstances where diagnostic AI is merely providing recommendations on diseases to a physician or radiologist, for whom the tool is advisory but not decisive, the level of explainability may be less necessary. Furthermore, in some breast cancer X-ray screening programs, where the AI tool is used to replace a human radiologist assessment, the approach currently adopted is that the technology needs to go through a complete clinical trial and appropriate regulatory approval. Equally, if the AI usage involves treatment recommendations, we would, in the midterm at least, always expect a clinician to make the final recommendation. So, even within one specific sector of healthcare, explainability requirements or alternative checks and balances will vary according to usage. To be accepted by practitioners and patients alike, AI tools will need to pass the existing high standards governing product liability and work within strong regulatory frameworks. Continuous revalidation of results will be paramount. Not all domains or use cases will demand this level of security. For example, as well as supporting customers in the medical sector, RELX uses AI tools in book or article recommendation offerings to researchers, and with our Lex Machina product we offer our legal customers insights into likely judicial outcomes in case reviews as a decision support tool. Whilst as with all our products, quality and accuracy is paramount, in terms of system explainability the risks of failure differ depending on what is being done, and as such the expectations as to transparency will vary accordingly. To ensure trustworthiness, each organisation will need to look at what data is being used, for what purpose, what analytic techniques and models are used, and what the results will deliver. It is an inherent feature of some AI "black box" technology that it is not readily comprehensible, but organisations	As AI is not a monolithic technology, guidelines and frameworks will need to be proportional, risk based, and flexible. Organisations should be encouraged to establish guiding principles, possibly through the mechanism of ethics boards. Businesses and policymakers need to continuously listen to concerns of stakeholders and build public understanding, confidence, and acceptance of AI developments. We would also encourage policymakers to cooperate at international level on ethical guidelines helping to ensure an inclusive and global approach.

RELX has a long history of developing products which combine content with powerful analytics. As technology advances, much of those analytics are based on natural language processing and increasingly on the use of machine learning to train the algorithms. We are already incorporating AI technology into many of our products. RELX group has always demanded and delivered many of the elements listed in the Guidelines as essential operational requirements. For example, requirements for good Data Governance and Justice are already written into all our business practices. This requirement exists independently of AI analytics. Similarly, the Design for All and Robustness requirements are already a given in our product development. AI will not change that.

Whilst there may be certain specific issues that arise directly from AI that have not needed to be addressed before, we would urge policymakers not to view AI as demanding reinvention of approaches and good practices that already existed in the pre-AI world of data analytics.
This also applies in other legal and policy areas. For example, in addition to providing knowledge to researchers, published scientific materials are now being used as training data for machine learning in developing AI products. Books, journals and databases are used to help to train machines to diagnose disease, predict weather patterns and develop new applications for drugs. The accuracy of the scientific record maintained by science and academic publishers is important to help ensure that machine learning has both depth and accuracy. This has historically been achieved by combining strong quality control mechanisms with robust intellectual property protections. The advent of new AI technologies must not be used as an excuse to weaken these established frameworks.

must be prepared to communicate key factors that go into the results and provide evidence of continuous testing and validation of models. It would not be appropriate to have to disclose source codes as it would not be useful for accountability purposes, could allow bad actors to avoid detection and would weaken economic incentives to invest development resources.

Human engagement and oversight

For AI to be trustworthy it will require citizens to comprehend in broad terms the basis of the decision making, the benefits of the system, and the checks and balances and redress systems in place for any perceived unfairness. Accountability, auditability and availability of human review will be heavily context and use case dependent.

Non-technical methods
As stated in the Guidelines, many regulations exist today that increase AI trustworthiness, and sector regulators will have an important role to play as the technology develops affecting their domain. Regulations should be reviewed and adapted as required. Attempts to horizontally regulate all AI activity would in our view be unnecessary, premature and inhibit innovation. The nature of AI makes it difficult to imagine a horizontal standard that would be meaningful across applications and sectors.

Maika	FOHRENBACH	American Chamber of Commerce to the EU (AmCham EU)	We welcome that the Commission's High-Level Expert Group (HLEG) on AI addresses the important topics of ethics and policy through an inclusive and multi-disciplinary approach. Moreover, we strongly agree with the objective of the HLEG to issue guidelines on ethics in AI (hereafter 'the Guidelines') that are actionable and proportionate and encourage companies to adopt a responsible and ethical approach to AI. This will be the real added value of the Guidelines and, whilst the current draft is a very good basis, it needs to be further 'operationalised' through the development of use cases. Finally, we believe that Europe is uniquely placed for the development of AI, and therefore agree with the intention to use these Guidelines to foster a reflection and discussion on the use of the technology at a global level.We support that 'Trustworthy AI' (p.1-3) has two components: it must have an ethical purpose and it must be technically robust. This will make sure that AI is trusted by its users and that it will result in improving Europe's competitiveness. Whilst regulation already exists that applies to AI – as rightly highlighted in the Guidelines – a common approach on ethical questions, principles and values brings a huge benefit in generating user trust and facilitating a broader uptake of AI. Furthermore, building trust is also a mean to demystify some of the scepticism around the technology. Trust is essential to create the needed dialogue for educating the public on what AI is and how it can be used. This could be clearly stated in the introductory section (p.1, 'Trustworthy AI').The Guidelines provide a thoughtful and comprehensive set of ethical considerations designed to help developers and implementers of AI achieve 'trustworthy AI'. In offering these considerations, the Guidelines acknowledge that 'different situations raise different challenges' (p. iii). We strongly endorse this point and believe that contextual considerations merit greater attention in the Guidelines. The degree of risk of individual or societal harm, and the potential severity of such harm, will vary enormously depending on the specific AI application at issue. In fact, many of the ethical issues identified in these Guidelines only arise for AI systems that have a consequential – or meaningful – impact on individuals. We therefore urge the HLEG to make clear at the outset of the Guidelines that their recommendations are not 'one-size-fits-all', and instead should be tailored to each specific implementation of AI depending on a careful and thorough risk assessment. We appreciate the target of putting a method in place to enable all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis, as this will help to support transparency for users and build trust. Considering the importance of the contextual elements – outlined above – only guidelines that are holistic, proportional, risk-based and flexible based on high-level principles, best practices, voluntary and industry-driven standards and existing regulation, seem appropriate for commitment to the proposal of the AI HLEG. Currently, the technical (and non-technical) methods (as mentioned in Chapter 2) can be outpaced too quickly by technology, and the detailed checklists (as described in Chapter 3) are both too specific and not relevant for all use cases. Thus, we would propose to not include these into the	We value highly the approach of the HLEG to derive the responsible and trustworthy development, deployment and use of AI from the fundamental rights, ethical principles and values that underpin the commitment of the European Union (EU). The purpose of AI should be to bring benefit to individuals, society and business, and we believe that a human-centric approach to its growth is the prerequisite for its lasting success.Taking into account that AI should create added value to different stakeholders, economic interests have to be considered as legitimate interests for a company in order to promote economic growth. Adhering to principles such as traceability, transparency and self-determination might come to an economic cost. Therefore, as an overall comment we recommend complementing the principle of 'beneficence' with a notion of proportionality in Chapter 1. Fundamental rights of human beings (p. 7): We appreciate that the commitment to preserve human rights and fundamental values is integrated into the AI ethics. Under the section 'human dignity', we would suggest to remove the opposition between individual and data subject. To be treated as data subject does not mean that human dignity is negatively impacted. Under the section 'Respect for democracy, justice and the rule of law', 'human-centric appeal, review and/or scrutiny of decisions made by AI systems' is not a right in itself but rather an 'opportunity' which exists as a consequence of the rights explained in this section. Ethical principles in the context of AI and correlating values (p.8): We generally support the five ethical principles and correlated values proposed by the HLEG (do good, do no harm, preserve human agency, be fair and operate transparently), to ensure that AI is developed in a human-centric manner. Whilst we agree with the principles in general, we must stress that in some cases the principles may be in conflict with one another and developers might be faced with contradictions. Therefore, we suggest creating a form of hierarchy within them and establish a resolution mechanism in cases of contradiction. The role of experts in this process is welcome. However, we would like to stress that there will need to be cross-disciplinary or legal experts depending on the issues. In addition, we encourage the HLEG to more explicitly recognise the fact that there will necessarily need to be a balancing between benefits and harms when deploying AI, and that some trade-offs may be unavoidable. Advancing the interests of certain individuals may inevitably impose harms on others (eg, an AI tool that makes one company more efficient might 'harm' rivals by making them relatively less able to compete).Moreover, we would like to add the following comments below: • The principle of beneficence: 'Do Good'We agree with the AI HLEG that AI should be applied only when an added value can be generated for people and emphasise that this added value can also be of economic nature, such as an increase of efficiency, accuracy, reliability or reproducibility. We also recommend that the Guidelines adopt a broad understanding of beneficence. In fact, AI can be a tool to improve wellbeing, preserve dignity and foster sustainability but it can also serve more neutral objectives whose direct individual or social benefits are less clear. We therefore acknowledge that AI	The Guidelines' conception of 'data governance' is too narrow and is not reflective of the fact that governance structures necessary to develop AI ethically include a broader range of engineering and design practices (eg, access controls, systems documentation, etc.). We therefore urge the HLEG to recognise that data governance is complex in practice and will need to be tailored to individual scenarios.In general, this chapter contains too many requirements and will make it challenging for developers to make them operational. For the sake of clarity, we would recommend amending and merging some as follows: • 'Data quality and governance' and 'respect for privacy';• 'Design for all' and 'non-discrimination';• 'Governance of AI Autonomy' and 'respect for human autonomy'; and• 'Robustness' and 'safety'. Requirements of Trustworthy AI (p.14)1. AccountabilityAs rightly written by the AI HLEG, the topic of accountability is highly dependable on the use case, the field of application, the autonomy of the AI and many more factors. A general approach or 'one-size-fits-all' solution should not be targeted. We would add the sentence, 'accountability might include the ability to contest the output and provide feedback on why a certain result is right/wrong', which is essential to learning systems. 2. Data Quality and GovernanceWe would amend the title as follows: 'Data Quality and Governance'. Data quality and data integrity potentially have a big impact on the AI-systems. It is important to be aware of how limited data sets, bias in data or other factors impacting data quality can directly affect AI-based recommendations. All stakeholders should aim at the reduction of unfair decisions (eg, due to bias), and increase transparency to continuously improve our data sets and AI systems. Furthermore, it needs to be recognised that in certain cases bias is intended because of the objective of the AI system. 3. Design for allWe support the observation that 'systems should be designed in a way that allows all citizens to use the products or services [...]' (p. 15), as our members' aim is for widespread adoption of AI technology in a way that is beneficial to society. At the same time, this requirement should also recognise that some flexibility may be needed when determining how to design for all, depending on the product or service concerned. 4. Governance of AI autonomy (human oversight)We fully agree with the analysis done by the AI HLEG that the concrete ways to implement human oversight (through safety, accuracy, adaptability, privacy and explicability) will differ depending on the application and specific AI systems. More concretely, we propose to always review what level of autonomy in decisions should be applied (ie, distinguish between AI used only as a source of information, AI as an assistant with final decision by user, or AI that acts fully automated without human involvement). Furthermore, we find it essential to also review the level of autonomy in learning (may the AI learn on the market (retraining possible), with limited parameters (no safety relevant parameters), or if no learning or evolvement on the market is possible). As a third dimension, we suggest reconsidering the level of risk (eg, which persons or laws could be harmed and how). On this basis, a use-case specific	We generally support the 10 requirements for 'Trustworthy AI', but believe that the long list of questions requires more work to be used by developers. In its current form, this chapter is frequently inconsistent and repetitive and the questions it seeks to answer are often too high-level to have any practical use.Furthermore, precise questions for AI auditing or assessing will vary from use case to use case, and a tailored response needs to be provided for each specific situation or question. The development of use cases will be essential to make the guidelines practical and actionable. We strongly encourage the HLEG to continue and focus the work on such use cases, in cooperation with industry and civil society. The following items are crucial:• Proportional, risk-based flexible and voluntary guidelines (not all questions are necessary to consider for a 'simple AI tool');• A 'holistic' approach with high-level questions to easily identify critical topics/use-cases, etc.; • Clear definitions (e,g, AI taxonomy, levels of transparency, bias); • Measurability of the questions; and• Clear differentiation of the questions and thus how it should be implemented, eg:o What is use case specific or has a broader context; o What is legal/ethical topics and what are technical solutions; ando Responsibility of business/development function and of governance function.	The American Chamber of Commerce to the European Union (AmCham EU) believes that key success factors for ethics in artificial intelligence (AI) guidelines are: • A human-centric framework that is proportional, risk-based and flexible because AI is not a monolithic technology but its ethical risk changes drastically according to its use and context. • A 'holistic' framework including high-level principles, best practices, voluntary and industry-driven standards and existing regulation. • A common understanding of what the problems are and further research to enable these problems to be more effectively addressed (eg, technology can also be part of the solution). • Encouraging companies to self-regulate: companies should establish guiding ethical principles for themselves that will apply throughout all their operations. • Encouraging companies to adopt concrete governance practices: AI ethics should be built into business performance, not bolted-on as an afterthought. AI ethics should be part of the AI lifecycle, from the data models and product deployment, to the update of workflows, tools, and business processes. For example, companies could set up internal structures such as 'AI ethics boards' to discuss these issues.• Encouraging companies to understand the key issues and tools to mitigate risk: as businesses and industry continue to pilot, adopt and rely on AI technologies to reshape the future of decision-making, AI that can be trusted to be transparent, fair, explainable and secure is imperative. Businesses need to continuously listen to concerns that might exist and adapt their ethical guidelines in developing tools to mitigate risks. • Continuous dialogue with stakeholders (industry, researchers, etc) on the development of appropriate mechanisms, in particular for any consideration of 'regulatory' mechanisms. • Increase overall awareness and foster trust through the entire value chain, from developers to users, as well as consumers and society at large. • Encouraging authorities to collaborate with industry and civil society in building data ecosystems which help to generate datasets in quantity and quality which ensure and empower a fair and ethical AI.• Encouraging policy-makers to cooperate at international level on ethical guidelines, helping to ensure an inclusive and global approach.

sections subject to formal endorsements. Finally, we share the view that the issue of AI ethics requires a regular discussion and iteration. Therefore, it would be helpful to clarify in the final version of the Guidelines the process that would be followed for this continuous dialogue, as well as regularly updating the document. Glossary (p.iv)In the definition of 'artificial intelligence (p.iv), we believe the final Guidelines should provide greater clarity on what is meant by 'deciding the best action(s) to take'. In particular, Article 22 of the GDPR articulates the concept of a 'decision based solely on automated processing'. Is the AI definition set forth in the Guidelines coextensive with the GDPR, or is it narrower (or broader)?In addition, we note that the Guidelines' definition of AI is narrower than many common understandings of the term. Many solutions in use today that are described as having an AI component do not necessarily 'decide' on a course of action; instead, many of them make connections, reveal correlations, or provide other insights that humans then use to decide on a course of action. We therefore believe that the proposed definition of AI should be modified to reflect this point. The Guidelines define 'bias' as 'prejudice for or against something or somebody, that may result in unfair decisions' (p. iv). In view of most data scientists, virtually any datasets will reflect at least some types of bias (eg, traffic data collected in large cities might not accurately reflect traffic patterns in smaller cities). The goal should not be to eliminate all biases in datasets used to train AI, as this is effectively impossible for most (and possibly all) finite datasets. Rather, the goals should be: (i) to take steps to mitigate the risk that an AI solution might generate unfair biases; and (ii) to help people understand the scope, characteristics and limitations of the dataset(s) on which an AI solution was trained, so that people can better understand how these limitations might impact the outputs generated by the AI in any given application. Finally, the Guidelines frequently use the terms 'transparency', 'explicability' and 'explainability' interchangeably. In our view, 'transparency' is a broader concept than 'explicability', the latter being also linked to an important separate term, 'intelligibility', that is somewhat overlooked in the Guidelines. We therefore encourage the HLEG to include each of these four terms in the glossary to help clarify intended meanings for stakeholders.

solutions may satisfy beneficence as long as they serve a useful purpose (to someone) that outweighs the risk and severity of potential harm to others. We fully agree that AI can help with societal issues, such as fairness and inclusion. We would welcome initiatives from the European Commission to foster the discussion and research on increasing the benefit of AI regarding ethical and socio-economic challenges. • The principle of non maleficence: 'Do no Harm'We agree that AI systems should protect the dignity, integrity, liberty, privacy, safety and security of human beings. AI applications are being developed by humans and it must be understood that high efforts and continuous improvements are necessary to reduce potential risks. Specific to each use case, the necessary quality level and fault tolerance, including fall back solutions in case of error and the required effort for testing, monitoring and controlling must always be defined under consideration of the field of application, such as: (i) how autonomous the AI may act (ie, differentiate between situations in its sole purpose as an information source, an assistant function in which the final decision is with the human, or if it is completely autonomous); (ii) how autonomous it may learn (ie, if re-training on the market possible and to what extent); (iii) the opportunities and possible risks and which machine learning method is used. • The principle of autonomy: 'Preserve Human Agency' AI can help people making better and more informed decisions. The two aspects highlighted in this paragraph are essential. First is the matter of choice: where possible, an alternative to being subject to direct or indirect AI decision-making should be provided to the user. However, it should also be considered that there are technological limits and that where real alternatives are possible today, there will be even more in the future. The right to opt out and withdraw from all AI decision-making does not seem realistic with the increasing use of the technology and will be difficult to implement. This might cause harm to others or prevent an authority from performing its duties for the common good. Such a 'right' cannot be horizontal – it must vary according to the use case and should be based on the type of AI system (the sensitivity of the use case).Second, is the matter of transparency. When an interaction with AI is taking place and where crucial decisions are made by algorithms, a risk-adequate and use case specific approach is crucial. However, in the current phase of the deployment of AI, we believe that it should be transparent where and in what form AI is being used. • The principle of justice: 'Be Fair' AI systems should be designed in a way that the predictions resulting from training data are fair and as unbiased as possible. Because AI systems are designed by human beings and are trained using data that reflects our imperfect world, it's important that developers are aware how bias can be introduced into AI systems and how it can affect AI-based recommendations. We should target the reduction of unfair decisions (due to bias) and increase transparency. At a minimum, AI-based solutions will increase consistency and deliver a standardised approach/decision. As noted in our comments on the definition of 'bias', removing all forms of bias from any finite might not be possible, which the draft decision should be taken. Moreover, a fallback solution for fully automated AI-systems should be prepared when human involvement is necessary. 5. Non-discriminationAs mentioned in our comments on the scope, most data scientists would agree that virtually any datasets will reflect at least some types of bias. Therefore, the Guidelines should not be aimed at eliminating all biases in datasets used to train AI, but rather at better understanding how limitations in datasets might impact the outputs of algorithms and taking all necessary steps to mitigate the risks these limitations might generate. 6. Respect for (and enhancement of) human autonomyWe agree that the user's well-being should be central to AI deployment and AI-systems should, where possible, promote conscious decisions by the users regarding the delegation of responsibility to the system. Applied well, AI could enable people to indicate much more precisely their preferences than would be otherwise practically feasible.7. Respect for privacyWe have a strong framework for data privacy in Europe with the GDPR and this is applicable to AI-systems. The Guidelines could mention the use of encryption, pseudonymisation and other privacy protective techniques that reduce privacy risk to individuals while still allowing for the development of AI solutions that can be beneficial to society. Moreover, they could stress that AI can be used to enhance privacy and its potential to do so should be further explored. However, no new framework or regulation is needed specifically to address AI. 8. RobustnessThe algorithms must be secure, reliable and robust enough to deal with errors or inconsistencies during the execution, deployment and user phase of the AI system. Therefore, there must be an extensive design and development phase, during which developers take appropriate measures to ensure safe and robust operation in the public. Reliability & reproducibility – We do not see these elements as core to the development of 'trustworthy AI'. The draft Guidelines should consider the limits of reproducibility, especially when placed on the market and retraining (by the user) is possible. These aspects should be handled by the overall system design and by extensive testing before and after being placed on the market. Accuracy – Accuracy of AI systems is limited and is directly linked to the data set used for training. More guidance is needed on what level of accuracy is required for AI systems, especially for sensitive use cases. Fall-back plan - The fall-back plan should depend on the use case and may not always be necessary.9. SafetyIt is worth adding that in many applications machine-learning increases the overall performance of a system, including in terms of safety compared to a strictly rule-based system.The AI system should be safe and not harm the user or his/her rights, it also should be reliable and do what is expected. Minimising the risks of the whole system, testing and quality monitoring will be key elements besides setting the right quality criteria (such as false positive vs. false negative rate). 10. TransparencyIt is important that there is a certain basic level of transparency or explainability to earn the user's trust. On the other hand, greater transparency will be necessary for

Guidelines themselves recognise (see p.16). Hence, we would encourage the HLEG to revise this principle to target 'unfair' bias. • The principle of explicability: 'Operate transparently'We fully agree with the AI HLEG that transparency and explainability are the key success factors to increase the acceptance and trust in AI systems. We agree that transparency means that the function of AI is explained an understandable manner, however, we would also add a contextual consideration in that the level of transparency depends on the application. In terms of 'business model transparency', we believe that the first basic requirement there is the need to inform individuals on whether or not they are interacting with an AI system. Beyond this, it means explaining the result, the base for decision-making and the benefit of the system. Providing the user with transparency, though, should not be afforded at the expense of a company's business model as this is sensitive information, impractical and often unachievable as they evolve frequently.We agree that explicability is a precondition of trust, however the draft guidelines seem to confuse general principles with AI-specific issues by linking explainability de facto with 'informed consent'. 'Informed consent' is a GDPR term with a specific meaning, and therefore its use with regards to explicability must be more precisely defined. As GDPR allows data processing based on legal bases other than consent, like legitimate interest, we suggest removing this concept altogether and replacing it with the focus of these guidelines – trust. Critical concerns raised by AI (p.11)• 5.1. Identification without consent:This chapter focuses on the consent in terms of privacy law. In addition to consent, the GDPR offers further legal grounds for data processing such as the contract and the legitimate interest. Therefore, the draft guidelines should not restrict the options that provide control to individuals as foreseen in data protection law.Regarding identification, we agree that there must be differentiation between the identification of an individual and the tracing and tracking of an individual, but one must also be mindful of what practices are harmful and not harmful, lawful and unlawful. Although we agree that the identification without consent could be a critical concern in some scenarios, it might not be a critical concern in others but actually beneficial. We therefore recommend that the final Guidelines approach this issue with a specific focus on use cases where identification without consent poses an elevated risk of harm to individuals or society. Moreover, the idea developed in the section of 'developing entirely new and practical means by which citizens can give verified consent to being automatically identified by AI or equivalent technologies', is a dangerous path towards a situation in which citizens' choices are overridden by others who think they made the wrong decision, simply because it is believed that they did not give it enough consideration. We also recommend that the Guidelines expressly acknowledge that different applications of AI might warrant different types of consent. In higher risk scenarios explicit consent might be appropriate, while in lower-risk scenarios, consent may be expressed implicitly, eg, by clearly informing a consumer that stepping into a store a store will entail the use of AI developers and operators to ensure quality monitoring and continuous improvement, as well as for use cases with potentially higher risks. Transparency levels provided should be contextualised and risk-based. However, achieving transparency can be complex and highly dependent on a host of variables, precluding anything resembling a 'one-size-fits-all' approach. Certain AI technologies, including deep neural networks, are so complex that they go well beyond what's comprehensible to humans. In these contexts, the overall goal of transparency would be ill-served. Stakeholders should be required to provide 'meaningful information' about choices and decisions concerning data sources and development processes and for uses that can have significant impact. Technical and non-technical methods to achieve trustworthy AI (p.18)The technical and non-technical methods listed to achieve 'Trustworthy AI' are a good non-exhaustive list, but more emphasis should be placed on the role of standardisation and codes of conduct. As the AI HLEG mentioned, the listed methods are meant to present only a sample of possible methods and should be continuously reviewed. They should only be considered as best practice and activities for further collaboration and research and therefore should not be included into the formal endorsements.Traceability & auditability – We believe that the development of human-machine interfaces that provide mechanisms for understanding the system's behaviour is essential. However, the nature of auditability will be heavily context-dependent. In complex scenarios, third party auditors and expert controls will be more effective for technical support. In still other scenarios, internal organisational auditing and controls may suffice. In light of this, the Guidelines should do more to acknowledge that effective auditing, depending on the context, can include any of those mechanics. Standardisation – We would like to stress that the nature of AI makes it difficult to imagine a horizontal standard that would be meaningful across applications and sectors.Accountability governance – The draft Guidelines rightly stress the importance of having a data governance programme with competence over AI. Whether this is specifically deemed an Ethical AI review board or whether it has a broader mandate also capturing AI, is perhaps less relevant and should depend on the scale and nature of AI work that a company is performing. When developing these mechanisms, the global dimension should not be forgotten. As mentioned in the executive summary of the draft Guidelines, we strongly agree with the view that AI and its development exists within a global ecosystem and therefore Europe should work tirelessly to shape the global debate on AI governance to promote trustworthy AI for all citizens.

					tracking to enable ‘frictionless’ shopping experiences. Finally, the Guidelines should note that many of these issues relating to identification – and so to processing of personal data – are already governed by the GDPR and other EU law. • 5.2. Covert AI systems:We agree with the statement that AI developers and deployers should ensure that humans are made aware of – or able to request and validate – the fact that they are interacting with an AI identity. In addition, we would note that the principle is potentially under and over-inclusive, depending on how one understands the notion of ‘interacting’ (p. 11).• 5.5. Potential longer-term concerns We suggest deleting this section. With the technology evolving, long-term impacts cannot be predicted. The probability of potential occurrences as mentioned by the HLEG (‘examples thereof are the development of Artificial Consciousness, ie, AI systems that may have a subjective experience of Artificial Moral Agents or of Unsupervised Recursively Self-Improving Artificial General Intelligence (AGI)’ p.13), are currently relatively low and well into the future. The purpose of the Guidelines is to be practical and immediately applicable by focusing on realistic and existing challenges while remaining attentive to future development of critical topics. There is no way to identify all possible scenarios, and we believe that the principles around which the Guidelines are built are broad enough to inform decisions on scenarios we do not foresee today. Ultimately, the goal of the Guidelines, and of any ethical principles in this space more generally, should be technology neutrality.		
Camille	Dornier	ACN - Alliance pour la Confiance Numérique	Artificial Intelligence (AI) will be fully part of our daily lives in the near future. Thus, these draft guidelines on ethics come at a perfect time. ACN truly believes that Europe should not lag behind in terms of innovation and competitiveness. As the High-Level Expert Group on AI (AI HLEG) rightly points out, Trustworthy AI can be -and should be- a competitive advantage for the European industry. “This is the path that we believe Europe should follow to position itself as a home and leader to cutting-edge, secure and ethical technology.” (page 1) ACN is totally in line with this statement, which accurately shows that ethics and competitiveness can go hand in hand.	ACN supports this “human-centric approach”, based on the core fundamental rights enshrined in the EU Treaties and the Charter of Fundamental Rights. The Paris Call for Trust and Security in Cyber Space (Appel de Paris) states that human rights offline must be protected online, and that international human rights law applies to the internet. ACN has signed the Paris Call and is fully committed to the protection of human rights. Principles derived from these human rights, such as Beneficence, Non-Maleficence, Autonomy of humans, Justice and Explicability should always be considered when developing, deploying and using AI.	This chapter mentions technical methods and non-technical methods to achieve Trustworthy AI. Among non-technical methods, the draft guidelines refer to standardisation because “[u]sing agreed standards […] can function as a quality management system for AI offering consumers, actors and governments the ability to recognise and reward ethical conduct through their purchasing decisions”. (page 21) Our association strongly believes that the uptake of European standards from CEN/CENELEC/ETSI or international standards from ISO/IEC is an efficient method to ensure that AI systems are in line with clear objectives. This facilitates comparison between such systems and thus brings more trust to the market. Likewise, the use of certification and AI assessment, based on standards, guarantee that products, services or processes are evaluated against a set of requirements. These requirements can relate to cybersecurity or privacy. For instance, Trustworthy AI can be achieved by certifying that an AI-enabled device is GDPR compliant. Compliance with EU Law (eIDAS, NIS, GDPR, Cyber Act etc.) is also a must and will provide to European industry a significant competitive advantage.	ACN would like to take the opportunity of this consultation to propose a few more assessment questions relating to the robustness of AI systems. ACN believes that this part could be further developed, as resilience to attacks is key to achieve Trustworthy AI. The following questions could be added to better assess resilience to attack: Is the AI system vulnerable to experimented and resourceful cyber attackers? Was penetration testing performed to guarantee that the system is cyber secure? Are there any backdoors in the AI system? Were the cybersecurity solutions used in the AI system certified? If so, against which standards and -existing or to develop- specifications?	ACN fully supports the document drafted by the AI HLEG. These guidelines could be further developed when it comes to assessing the robustness of AI systems. The number and diversity of cyber attacks is increasing every year and, if AI is to take a central place in our daily lives, it is of utmost importance to ensure that AI systems are resilient to attacks. Resilience to attacks, as well as conformity assessment, is per se an essential requirement to achieve Trustworthy AI. Moreover, resilience to attacks is also an absolute precondition to fulfil other requirements, such as the respect for privacy. No privacy will ever be achieved if an AI system is vulnerable to attacks.

Anonymous	Anonymous	Anonymous	Mastercard is grateful for the opportunity to contribute to the consultation on the European Commission's High-Level Expert Group on Artificial Intelligence ("HLEG for AI") Draft Ethics Guidelines for Trustworthy AI. We are passionate about technology and are pleased to share our views in respect of the European Commission's strategic vision for a trustworthy AI as part of the future of the European digital economy. Mastercard is a global technology company in the payments industry. We operate the world's fastest payment processing network, connecting consumers, financial institutions, merchants, governments and businesses in more than 210 countries and territories. Our products and solutions make everyday commerce activities – such as shopping, travelling, running a business and managing finances – easier, more secure and more efficient for everyone. Mastercard is committed to ethical use of data and robust standards of privacy and data protection. We believe that innovation and individuals' rights go hand in hand, and that the best outcome is achieved when innovation is balanced with responsible and ethical use of emerging technologies. For this reason, we place the individual at the centre of everything we do and invest considerable resources in human-centric innovation. Executive GuidanceWith respect to the reference to Chapter I: Key Guidance for Ensuring Ethical Purpose: "Acknowledge and be aware of the fact that, while bringing substantive benefits to individuals and society, AI can also have a negative impact" in the Executive Guidance of the Draft Guidelines: We acknowledge that AI technology bring substantive benefits but may also have negative impacts that need to be balanced against each other. To that end, the following considerations can be made: 1. Individual impact – Many AI applications even though important for the society may have very limited (or no) impact on individuals. For instance, AI models on trends in manufacturing industry. The Guidelines should make it clear that a higher level of scrutiny is required when AI applications affect individuals, and should introduce a risk-based Approach. To illustrate, AI technology is likely to have a higher impact (or risk) if it is ultimately applied to an individual as opposed to when the AI technology is used to generate aggregate insights. For instance, a bank developing an AI model on aggregate preferences between credit or pre-paid cards can steer the bank's investment funds towards the most profitable business strategy. If that AI model is applied to how each individual interacts with the bank separately, the impact is likely to be higher, as the bank can potentially use the insights from the AI technology to adjust its advertising campaign or marketing strategy to that particular individual. 2. Reticence - In addition, because AI may have a positive impact on the individual and the society, as mentioned in the Executive Guidance of the Draft Guidelines, reticence risks, i.e. the risk of not using AI which would deprive the society from great benefits need to be taken into account. For example, in recent years, the payments industry has benefited significantly from AI technology in the fight against fraud. As fraudsters become smarter and smarter, a potential reticence risk in using AI technologies in detecting and	The principle of Beneficence: "Do good"We agree that AI systems should be developed to improve individual and collective well-being. However, we would like to highlight that sometimes, the individual and collective well-being may be in conflict. In those cases, there needs to be a balance between collective and individual well-being stemming from the use of AI. Beneficence for the society may have negative impact on some individuals. For example, profiling a fraudster who tried to impersonate a legitimate credit card holder may harm the individual fraudster but benefits the legitimate card holder and the society as a whole. The Principle of Justice: "Be Fair"The principle of fairness is a cornerstone in developing and using AI technology. Whereas we agree that all actors in the supply chain need to ensure that individuals and vulnerable groups maintain freedom from bias and discrimination, we want to emphasize that bias and discrimination in human decisions existed prior to AI technology. Therefore, AI technology should not be subject to standards that are impossible to comply with even in the existing environment. There are statistical tools used to monitor bias today. Companies should be able to continue to rely on those tools, as well as improve on them using AI.Rather, it is important to recognise that AI can assist in identifying and addressing bias and discrimination that exist today. This can be achieved through accountability practices such as audit, regular evaluations, documentation, data governance and other remediation mechanisms that may reveal instances of bias in the AI technology. The principle of Explicability: "Operate transparently"Transparency is a key pillar in maintaining citizens' trust in AI technology. Given that AI technology can be very complex, the same level of transparency may not be appropriate or achievable for all types of audiences – or for all uses of AI.For example, full technical transparency in the context of fraud detection and prevention would not be appropriate as it would give more ammunition for fraudsters to circumvent the fraud detection and prevention AI solution. At the same time, providing technical transparency to the regulators in case of investigation is more appropriate given that they would have an appropriate background and resources to review the technology. For some uses of AI, transparency might be better served by explaining the decisions, and not the model. For example, explaining to individuals that a company is using AI to determine to handle customer service inquiries and the data used to do so would likely be clearer – and therefore provide more transparency – than providing dissertations on natural language processing.In those cases that is not achievable or appropriate to ensure full transparency, we recommend considering how to compensate for lack of full transparency through other means. This could include human review of AI decisions, redress mechanisms and ensuring detailed technical transparency to regulators in cases of investigation. We suggest the consideration above is also reflected in the Assessment List (Section III) by adding a question on adjusting the right level of transparency to the audience.Critical concerns raised by AI5.1 Identification without consent & 5.3 Normative and Mass		We agree with the approach to operationalize the implementation and assessment of requirements. However, because of the variety of contexts that AI technology is used, it may be difficult to ensure that all questions apply to all contexts and that thes e are exhaustive.In addition, when the GDPR is referenced in section "Respect to Privacy", we would welcome a clarification that the GDPR would govern the processing of personal data i.e. the processing activities/data uses performed by AI systems, not compliance of AI systems as such.	Mastercard would like to thank the HLEG on AI for receiving our contribution to this important consultation. As a global technology company, we are committed to privacy and security and the responsible and ethical use of data. We are also committed to working with the HLEG to craft a framework for ethical use of technology in a manner that protects the individual, ensures ease of commerce, and helps to further develop and strengthen digital innovation in Europe. We remain at your disposal for any further assistance you may require.

			preventing fraud would have as a consequence high amount of fraudulent transactions in the e-commerce space.	Citizen Scoring without consent in deviation of Fundamental RightsArticle 6 GDPR provides a number of legal grounds which may be leveraged depending on the use case of AI technology. The reference to consent in the title of 5.1 and 5.3 may create an impression that consent is the only legal ground that could be acceptable in AI technology applications. We would therefore like to highlight that the legal ground for processing personal data under the GDPR may be selected depending on the specificities of each AI technology application. Consent may be one option but a legal obligation or the legitimate interest of the data controller or other parties may be other. For instance, the revised Payment Services Directive (EU) 2015/2366 ("PSD2") includes a legal obligation for the payment service providers to perform transaction risk monitoring and analysis in order to ensure security in remote payments. AI technology may be used to assist transaction risk monitoring and analysis as a means to comply with the legal requirement from the PSD2. In this case, the processing of personal data is necessary for compliance with a legal obligation that the data controller is subject to, whereas consent is not workable in practice and not optimal (as Fraudsters would never consent in the first place). In addition, it should be noted that the GDPR is technology neutral. Therefore, a legal ground for processing must be chosen based on the type of processing activities performed on personal data and not because of the use of AI technology as such. Accordingly, we would welcome clarification that there is a need to secure a legal basis before processing personal information irrespective of the technology used, and that consent is only one option amongst others according to the GDPR. While consent is desirable in certain cases (e.g. collection of sensitive data), it is not in others (e.g. transaction risk monitoring required by PSD2).The questions seem to focus on system compliance with the GDPR. We would welcome a clarification that the GDPR would govern the processing of personal data i.e. the processing activities/data uses performed by AI systems, not compliance of AI systems as such.		
Mariana	Damova	Mozaika	My main remark to this extensive and comprehensive document pertains to the definition of AI, that lays the basic assumptions for the content of the document. Having consulted several sources and relying on my professional experience, I would suggest the following definition: "AI means algorithm-based and data-driven computer systems that enable machines to perform human like actions, such as learning, acquiring and processing information, reasoning, decision making, adapting, self-correcting, visual perception, speech recognition, language understanding and translation between languages. They use human reasoning as a guide to provide better services or create better products through digitalization and partially in collaboration with humans." The above definition includes several points that have not been mentioned in the document's definition, such as human like	Section: Fundamental Rights of Human Beings Along with the enumerated rights, I would suggest a word on resonsibility, that will lead us to explainable and responsible AI, as part of the Thrustworthy AI subsection: "Do no Harm" Do no harm to the human beings and individuals, but also to the nature, the environment, the economy, any economic activity, the society may be considered as broader scope of the statement Vulnerable demographics has an "e.g." list explaining who is meant as vulnerable demographics. In my opinion "immigrants" do not belong in the list, as they are people as everyone else. Subsection "be fair": Here "be fair" is projected to the regulations of AI. But AI has to be designed to be fair in its actions and decisions. The definition of fairness here	Subsection: Data Governance As stated above, AI should be able to detect and handle bias. This task is beyond single or collective human effort, unless we speak about wrong doing on behalf of the developers and the data operators. Subsection: Governance of AI Autonomy (Human oversight) AI to be by design ethical and autonomous. The Human oversight should be considered from this early stage on. Section: Architectures for Thrustworthy AI The authors speak about rules to control the actions of the AI. How about linking to the previous section X-by-Design, and consider building Thrustworthy AI by design. Many of the points in this Chapter can be conveyed in the points of the first Chapter so that the document becomes more consistent. It seems to me that there is a mismatch in the use of certain words, such	Section: Accountability I would suggest a point: check whether the AI behaves responsibly itself	Thanks for the opportunity to comment on this extensive and comprehensive draft document. My two main points regarding this document are that 1) the definition of AI as it is given excludes several facets of AI, that inevitably are talked about in the subsequent sections, and in my opinion it should be revisited. 2) when speaking of human centric ethical AI, the document focuses almost exclusively on the direct impact on humans as social beings, but maybe examples beyond the humans as social beings can be considered to make the picture of ethical AI more comprehensive.

			behaviour, human intelligence, information, natural language, speech, translation, adaptiveness. Related to the bias definition: the definition in the document states that bias can be injected by training data. This implies machine learning methods. My comment here is that AI is not only based on machine learning, so maybe the wording of this sentence can be made more generic in the direction of "importance of data quality and analysis for AI systems". Further, AI systems should be able to detect bias, and this in my opinion should be mentioned in the guidelines. Section: The role of ethics: When we think of ethical guidelines for human centric AI, we should keep in mind that AI systems will become autonomous, and the ethics rules have to be modeled and installed in AI's reasoning and decision-making processes. So, "the outcomes of the AI systems must be ethical" with everything that ethical implies: do good, do right, do no harm, include everyone, etc. may be added	should not imply that from the results of AI there will be negative impacts for somebody and positive impacts for somebody else. Be fair is treated narrowly only with respect to different social groups. AI is about decision making in a vast variety of domains. Even in a factory production setting, in IoT context, AI must be ethical and fair in its actions and decision making, and in this setting humans are only indirectly affected. The AI of autonomous cars must be ethical in the decision making of when to stop and when to turn left or right in certains situations. Subsection explainable AI: the term explainable AI is used to describe the field that tries to understand the decision making process of AI to make this process explicit. The section emphasizes the role of developers to instruct correctly the machines. I think the emphasis should be on the ability to explain how a conclusion has been drawn by the machine. If what I just described in included in the term "auditable", I agree with this. Section: Potential longer-term concerns The points raised in this section should be a starting point for building guidelines for Ethical AI to begin with.	as "explainable" in the first chapter and in the second chapter. To the invitation of suggesting a point to lead towards achieving trustworthy AI, I would suggest a point: AI, Trustworthy by design		
Luca	Cassetti	Ecommerce Europe	Ecommerce Europe is the voice of the European digital commerce sector. Through its 19 national e-commerce associations, Ecommerce Europe represents more than 75,000 companies selling goods and services online to consumers in Europe. Ecommerce Europe believes that the uptake of Artificial Intelligence Systems can have a major impact on business and citizens globally, becoming a key driver of economic development. In recent years, AI has been increasingly implemented into the consumer's e-commerce experience.The strategy adopted by the European Union is moving in the right direction, promoting technological developments to ensure that European companies can compete globally, while taking into account all the ethical and social aspects linked to such technologies. Ecommerce Europe, as a European ecommerce association, would like to contribute to this consultation by submitting its feedback on the Ethics Guidelines for consideration.Ecommerce Europe agrees with the observations raised by the European Commission and the High-Level Expert Group on Artificial Intelligence (AI HLEG) and welcomes the structure of the guidelines. The three pillars highlighted in the document will allow all stakeholders to examine in great details the crucial aspects of the uptake of AI. Indeed, when developing new technologies using Artificial intelligence and regulating these technologies, it is fundamental to keep in mind the rapidly changing social, economic and technological context to question and reconsider periodically our approach in light of technological developments which makes predicting different scenarios almost impossible. The complexity of this issue requires a systematic approach as it involves aspects that are profoundly different but inevitably interconnected, such as technology, ethics, regulation, economy, governance, etc. From this assumption derives also the difficulty of interacting with	AI can help create highly personalized shopping experience for consumers. Based on consumer data it receives (past purchases, buyer profile etc.), AI can offer consumers exactly what they are looking for. However, personalization will only be possible if consumers provide their data. The more data AI gets the more accurate it will be. At the same time, when it comes to personal data, there are some challenges. Namely, any development of AI technology will have to take into account data and privacy protection laws, such as GDPR, ensuring that AI complies with the strict regulation.Ecommerce Europe fully agrees with all the remarks developed by the A.I. HLEG in terms of "Fundamental Rights of Human Beings" and "Ethical Principles in the Context of AI". Defining the framework in terms of set of principles, values and purposes is crucial, as well as the possible risks that could arise long-term. This is the key to mitigate any negative effects while benefiting from the positive ones.Ecommerce Europe fully supports the principles set out in "Ethical Principles in the Context of AI and Correlating Values" (Chapter I, Section 4).Referring to Chapter I), Section 5, "Critical concerns raised by A.I.", Ecommerce Europe asserts that the most important points of attention have been highlighted by the AI HLEG. However, Ecommerce Europe would like to submit additional considerations.First, Ecommerce Europe observes that some aspects concerning data protection as a right of human beings have been correctly taken into consideration and analyzed (Ref. to (5.1.), (5.2.), (5.3.).The EU Treaties and the Charter of Fundamental Rights clearly establish these principles, followed by legislation such as the General Data Protection Regulation (Reg. 679/2016 or GDPR). In particular, the GDPR sets rules that define the perimeter and the limit of data treatment, but it does not explicitly define A.I.The Regulation mentions	Users should be confident when using A.I. and user trust in A.I. should be further reinforced. However, over-regulating could hamper the development of A.I. in Europe, leading to Europe's dependence on A.I. developed outside the EU. This non-EU A.I. technology may not necessarily be in line with the ethics guidelines the EU will have developed. This area is explained in more detail under Chapter 3.	In light of the above considerations, Ecommerce Europe submits its point of view about the Assessment List.Accountability: no comments to add.Data governance:Ecommerce Europe agrees with the remarks but would like to underline the need to better define the processes, the criteria for stating (human) responsibility on all levels of the treatment (i.e. both the development and use of algorithms) and the criteria for the valuation of the legitimacy for the use of data; further questions that could be raised, such as:"How could we determine the Jurisdiction in case of damage caused by incorrect data handling? Should the GDPR criteria be used?";"Should the Jurisdiction be established based on territorial criteria or matter?"; "In case of damages produced by machine learning to Third Parties (such as users), should the liability be imputed to the developer, to the owner of the machine, to the client, etc.?".Design for all: no observations about it.Governing AI autonomy:Ecommerce Europe fully agrees with all the assessments listed and considers the questions of governing AI autonomy a key issue; Ecommerce Europe adds that it could be helpful, in certain circumstances (especially where autonomous robots are allowed to interact in delicate areas) to set an internal supervisory mechanism whose members are external, nominated periodically, composed by technicians and experts from other fields (such as lawyer etc.) in order to value not only the first level of the process (such as the algorithms) but also the results, i.e. the output that could be generated by the machine (which it could not therefore be previously determined). Non-discrimination: no comments to add.Respect for Privacy:Ecommerce Europe supports and agrees with all the points; it adds that data protection is part of human rights and for this reason its protection should be applied independently from GDPR, being an instrument to implement this protection (it means that evaluating simply	Finally, following the recommendation of the HLEG to share comments about one of the areas mentioned in the documents, we close this contribution with some last considerations about the impact of A.I. on e-Commerce and Digital Marketing.Ecommerce Europe believes that systems based on A.I. can have a positive impact on the ecommerce industry, for businesses and users. They do however open new important issues.A.I. could be decisive to cut down on some important challenges encountered by online merchants, such as fraud detection and prevention for online transactions, to detect counterfeiting of goods, especially in the pharmaceutical or childcare sector, as well as identity theft, etc. Technologies like blockchain will allow users to have more control of their data thanks to greater transparency in management, the power to control flows, protection from threats and malware, as well as from the risk that data ends up on the deep web.A.I. could improve services and user experience in terms of search, features and personalization, contributing to recommendations and purchase predictions tailored for the users, or to the development of a predictive customer service. We should however be aware that unpredictable consequences can arise. We are already witnessing first cases where the systems based on blockchain technology are used to record, in a more effective and precise way, the customer journey along the entire consumer experience.The system becomes the collector of data across all touch points: this means that all user's actions and interactions are acquired and stored: from the opening of the email and newsletter, to the registration on a website and the following accesses or app downloads; from the purchase of the product online or offline to payment, the use of discounts and coupons, etc.All these actions are recorded on the ledger, validated with a certain date and made unchangeable, thus attributing certainty to the identification

the relevant Stakeholders for each sector to satisfy all needs.

automated decisions and profiling systems (Articles 4 and 22 GDPR), that are inevitably linked to the development of AI. It is important to consider that unpredictable developments could lead to consequences that cannot be fully assessed now. Therefore, principles and rules on treatment and profiling should be integrated. Following these premises, please consider the following:Referring to (5.1.):Through connected devices (such as smartphones) and services, companies are already able to collect data in every industry. In particular, the most advanced digital marketing systems process huge quantity of information every minute, often without making the interested parties aware.Another critical area where a huge quantity of data is generated and processed is the transport sector: for example, the rapid increase of sensors and cameras allow the massive collection of data. Related to these issues, in addition to the findings raised by the Group, Ecommerce Europe highlights other aspects that must be considered as it is increasingly difficult to determine the criteria for attributing ownership and legitimizing the use of data and related responsibilities. Referring to (5.2.):Ecommerce Europe notes that Citizen Scoring activities could become more pervasive with evident risks for fundamental rights, especially for some categories such as vulnerable people (minorities or disabled persons) or in situations where there are clear asymmetries of power. Relating to this point, it should be noted that even Article 22) of GDPR does not seem to cover all the possibilities and implications that could arise. The Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 proposes important leads to face these issues, but those are not enough to address all possible scenarios.   We underline that profiling processes and related services are becoming more and more widespread and are required both by companies and the Public Administration. Profiling activities and information about a customer have become a fundamental starting point for determining the success of companies’ business strategy. Concerning the Public Administration, a potential scenario could be profiling and predicting election results. Ecommerce Europe recommends investigating these aspects further.Referring to (5.5.): Ecommerce Europe believes that “Potential longer-term concern” refers to the point of Responsibilities. The A.I. Systems mentioned by the AI HLEG – (i) AI systems that may have a subjective experience; ii) Artificial Moral Agents; and iii) Unsupervised Recursively Self-Improving Artificial General Intelligence (AGI)) – could raise some of the most critical aspects in terms of civil and criminal liability of algorithms and advanced machines as well as jurisdiction issues.Until now, the direct responsibility of the machine has been rejected – considering that in most cases it was attributed to the producer and, in limited cases, to the developer – nonetheless, for future developments it cannot be excluded that the evolution of this sector will lead to previously unexplored legal issues.In particular, the following issues can be raised: i) How to qualify automatic systems from a legal perspective; ii) How to attribute responsibility if an AI (or machine learning) system causes damage

compliance with it seems reductive with respect to the values discussed). The developers should place the protection of human dignity in the center of their reflections. In any case, GDPR recognizes fundamental rights for users, it is therefore necessary to guarantee the respect of these rights even when the treatment is carried out by machines.Respect for (& Enhancement of) Human Autonomy:Ecommerce Europe totally agrees and adds some considerations. It could be useful, in certain circumstances, to set out verification procedures that could reply to questions such as: “Are there mechanism or systems that allow the user to easily submit a complaint to the owner of the process/ the machine/algorithm decisions in case of (alleged) prejudice?”“Is the machine/algorithm decision binding for the users?” “What are the instruments (included organisms) that are authorized to analyze the issue?”Robustness: no comments to add.Safety: no comments to add.Transparency:Ecommerce Europe totally agrees with these remarks and would underline the need to guarantee transparency of the processes, especially in certain sectors or areas where the human being could be prejudiced. Processes should be scalable and verifiable on all levels by independent and third-party assessment systems. Therefore, transparency should include: Processes; Liability for each level of the process; Assessment; Traceability of the data and the output; Mechanism to withdraw date, information and destruction of the most sensitive information.

of the person, his actions and his preferences. These systems can already react to events in the real world and are able to grasp - thanks to the acquired information assets - not only rational behaviors and interactions, but also irrational and unconscious behaviors. We cannot, therefore, exclude that soon they will also be able to guide them. The most critical aspects lay down on this point. The questions might, moreover, be: “To what degree is it possible to regulate these phenomena?”, “How far does the algorithm (and the learning machines) could be push forward?”, “How far can predictions and direct and indirect conditioning of human actions and thoughts be pushed forward?”. The questions are many. Moreover, although our reflections have been developed from a business strategy point of view, we need to ask ourselves what implications these developments might have if used by the public sector.Ecommerce Europe was pleased to contribute to this consultation. We remain available to support the Group of Experts in the development of their reflections.

				not planned by the user and not foreseen by the developer? iii) Is the machine imputable under criminal law? iv) Is there any responsibility for the developer even in the event of an error, defect or malfunction of an intelligent robot? More questions could be raised. Another aspect is related to Jurisdiction. The first question that must be asked is: "How to determine the Court jurisdiction in the case of damages determined by machine learning?"; "Could the current criteria be used?".Furthermore, additional legal and economic aspects could be mentioned; although they do not generate serious and direct consequences for human beings, they could nevertheless determine legal issues of considerable economic value.Ecommerce Europe refers to the issues related to Intellectual Property. Until now, only the aspects related to the development of computers and algorithms have been considered but, in the future, new and unexplored issues could arise, especially from a legal perspective. It means that it will be necessary to reflect on how to qualify and resolve possible controversies that may arise in relation to the intellectual works produced directly by machines. It will be necessary to give answers to questions such as: "Are the outcomes produced by robots or learning machines protected?", or "Can a project developed by a machine learning be patented?", and "How can we determine the territorial jurisdiction?".	
Wolfgang	Schröder	University of Wuerzburg			I fully subscribe to, and strongly support, the human rights-based approach and democratic spirit of the Guidelines. However, I would like to critically question the – in my view – structurally rather underdeveloped concept of what AI ethics is about that seems to inspire the argumentation oft he paper. Which ethics? Whose responsibility? Doing "AI ethics" basically means searching for, and - if possible - finding, plausible answers to these core questions. There can be normative guiding principles for this evolving new kind of research programs and AI-related ethical discourses: e.g. the human-centric imperative to strive for the best possible mitigation of moral ambivalence both in the development and application of AI-based agencies. Yet realistically, these moral principles for dealing with AI should not be understood as required for a strictly general "structure" or "field" of AI ethics. Rather, they should be designed for a plurality of different "area ethics" related to AI aspects. What is needed, instead, is a clear distinction and exact consideration of the different contexts and goals in which, or for which, AI systems are developed and applied. Overall, area-specific ethics requirements can be identified in at least four fields of digitization: (1) Software developers need a Code of Conduct that makes ethical principles

							mandatory when programming algorithmic decision-making systems; (2) For software providers, an ethically oriented Corporate Digital Responsibility strategy is needed, in addition to AI development criteria in the narrower sense. The quoted strategy should place the selection of possible AI-related business models as well as goals and fields of application of the products and services containing AI elements to be produced resp. delivered under a shared normative standard. (3) Societal debate requires an adequately moral science-based reflection on broadly consensual ethical principles that determine whether, and to what extent, AI programming, application, and evolutionary forms of AI-based agency can and must be public-interest-oriented. (4) Collectively binding decisions at the level of politics and law need an AI area ethics of their own. They require a democratically generated and legitimized "operative agreement" on which normative claims and principles should be guiding and binding for the development, marketing, and application of AI-systems. Therefore my suggestion to the High Level Commission would be to adequately consider this plurality of different area ethics that ultimately "amount" to "AI ethics" in the revision of its impressive document.
Eline	Chivot	Center for Data Innovation	The draft guidelines begin with an introduction that includes the definitions of key terms, including artificial intelligence (AI), ethical purpose, bias, trustworthy AI, and human centricity. It also recalls the process and the purpose of the HLEG, the intent of the consultation, the role of ethics in AI, and the scope of the guidelines. The HLEG should update and revise some of those definitions which fall short and clarify the implications for any stakeholders who may choose not to endorse the voluntary guidelines. The definition of AI and bias in the glossary merits further elaboration. In particular, AI is not entirely, as stated by the draft guidelines, "designed by humans." Some forms of AI, particularly those using machine learning and deep learning, build models from data that require little to no manually engineered intervention. Indeed, a goal of many companies is to construct machine learning systems that can build other machine learning systems, such as Google's AutoML. In addition, the guidelines' definition of AI specifically does not make the distinction between two very different types of AI: narrow and strong. Narrow AI, also known as weak AI, refers to machine intelligence able to perform a specific narrow task for which they have been programmed, such Apple's Siri virtual assistant, which interprets voice commands. Strong AI, also referred to as artificial general intelligence (AGI), is a hypothetical type of AI that can meet or exceed human-level intelligence and apply this problem-solving ability to any type of problem. The draft guidelines note that a "mechanism will be put in place that enables all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis." However, the draft guidelines contain no further information about the nature of this mechanism and what would be the	The first chapter lists selected fundamental rights, principles, and values which, according to the HLEG, AI should comply with to ensure its "ethical purpose" and trustworthiness. For instance, the fundamental right "respect for human dignity" leads to the "principle of autonomy," which reflects the freedom of individuals to make their own choices and is operationalized by the value of "informed consent." The chapter concludes with a section on "critical concerns" raised by certain uses, applications or contexts of AI, such as citizen scoring and Lethal Autonomous Weapons Systems (LAWS). This chapter contains multiple examples of a negative tone, flawed references, vague statements, and unrealistic requirements. First, accusing AI systems and industry of, for example, working against democratic processes and values, limits this document's legitimacy and credibility. The HLEG's statements about AI should be fair and balanced, and clearly distinguish when it is referencing speculative concerns versus proven ones. In addition, to understand the potential tradeoffs of limiting or slowing the advancement of AI, the HLEG should include examples of how AI solves many economic and societal challenges. Second, several instances in the guidelines suggest AI systems should be held responsible for achieving complete equality—an unreasonable standard that does not exist for non-AI systems and processes. The HLEG should also revise and clarify other unrealistic constraints and impracticalities, such as references to "high standards of accountability" which, left undefined, could lead to confusion and stifle innovation in Europe. Finally, concerns raised in the final section of this chapter with respect to explainability do not sufficiently credit the	The second chapter of the draft guidelines attempts to map the general principles of the first chapter into concrete requirements for the development and use of AI systems, and suggest a number of technical and non-technical methods to this purpose. But there are several problems with this section. First, some of the requirements to embed ethics within the design and development of AI systems would be unnecessary and counterproductive. Contrary to what the guidelines suggest, the developers and designers of AI applications cannot always be held responsible for ensuring equality and equity in the use of their technologies. AI is a multipurpose tool, and the ones who should be responsible for ensuring its appropriate use are the operators who deploy the technology. Should there be any oversight, it should be built around algorithmic accountability—the principle that an algorithmic system should employ a variety of controls to ensure the operator can verify algorithms work in accordance with its intentions and identify and rectify harmful outcomes. Second, requirements to have humans review certain algorithmic decisions raise the labor costs of using sophisticated AI systems which offer better accuracy. As a result, a right to human review of algorithmic decisions will force companies to use less accurate AI systems that may actually increase bias. Due process and scrutiny should always be appropriate to the nature and seriousness of the decision at hand, and not be based on whether the decision was made by a human or an algorithm.Third, the guidelines' methods recommend relying on human decisions to solve the "limitations" and "biases" of AI. This incorrectly portrays AI as inherently biased and human ones as unbiased. Yet human decisions are often less accurate,	The third chapter provides a list of questions to guide developers when designing AI systems, and to help them assess whether these comply with the requirements and ethical principles of "trustworthy AI." The use cases that will illustrate how this would work in practice will be provided in the next iteration of the guidelines and will be helpful to evaluate whether these questions make sense.Based on the comments and observations offered for the previous chapters of the guidelines, several questions could be refined or deleted.For the requirement "Accountability," the first question "Who is accountable if things go wrong?" is too broad and points to AI as holding intrinsic risks. Why not ask "Who is accountable if things go right?" Moreover, "wrong" is Manichean language that is not adapted to the way businesses make decisions, measure risk, and assess results. The guidelines encourage organizations to consider "diversity and inclusiveness" policies when recruiting staff working on AI. This is an important element. In many EU countries, such policies are compulsory, but not always efficiently implemented. Yet given the skills available in Europe may not match the demand and the needs for the development of AI and diversity, this may not always be practical for a business. Therefore, it cannot be yet a reliable measure of accountability, and this question should be positioned under another requirement, such as "Non-discrimination" or "Design for all."The guidelines ask "Has an Ethical AI review board been established?" While some companies may choose to use review boards, there is no evidence that this should be a standard. Moreover, this framing suggests that organizations can and should put in separate accountability mechanisms for uses of AI as opposed to other	The Center for Data Innovation is pleased to submit feedback to the High-Level Expert Group (HLEG) on AI on its draft AI Ethics Guidelines for Trustworthy AI. The Center acknowledges that this initiative is timely and supports it for having involved a broad diversity of stakeholders within the HLEG, and for its non-legally binding nature. The guidelines are an opportunity to further the conversation on AI, which given the stakes, is much needed to provide a sense of urgency to the European policymakers, business community, academics, and the general public about the potential opportunities to use AI to improve the economy and society. The emphasis on addressing the needs of vulnerable groups, ensuring diversity and inclusion, and addressing skills are key contributions in the HLEG's document.The guidelines aim to provide concrete guidance on how to implement and operationalize "trustworthy AI" systems that "maximize the benefits of AI while minimizing its risks." While this goal is worthwhile, the guidelines have four main problems: 1) they present an overall negative tone towards AI; 2) they overlook the importance of EU leadership on AI adoption as a means of influencing global AI ethics; 3) they incorrectly suggest that developing a European AI ethics governance system will allow the EU to significantly differentiate its AI solutions, thereby gaining global market share; 4) they inaccurately frame AI as a technology that requires ethical tradeoffs, instead of one that can be used to improve ethical behavior; and 5) they propose principles such as transparency and explainability that would limit AI development. First, when the HLEG recognizes that "on the whole, AI's benefits outweigh its risks," it is damning with faint praise. In fact, the overall narrative it

consequences for stakeholders who do not wish to "formally endorse" the guidelines. There is a risk that these voluntary guidelines may become an attempt at backdoor regulation, such as penalizing companies who do not adhere to it. To avoid that, the HLEG should not endorse any particular mechanism for stakeholders to adopt the guidance, but instead put it forth and let it stand on its own merits.

vast amount of research taking place to improve AI explainability. Other references even seem to discourage the integration and use of new technologies such as facial recognition, and fail to acknowledge the strategic importance of developing autonomous systems. This will limit Europe's competitiveness and its ability to protect its infrastructure while China and the United States, for instance, will be catching up and gain a competitive edge. Chapter 1, Section 3.1 ("Respect for human dignity") suggests that businesses developing AI systems would treat people "merely as data subjects" and not with dignity or respect. This accusation does not accurately reflect how businesses using AI treat their customers and look to AI to improve product and service quality. Indeed, many businesses are investing in AI to deliver better quality or value to their customers. Accusing industry of such attitudes further feeds into the false narrative throughout the document that AI is negative. Chapter 1, Section 3.2 ("Freedom of the individual") states that protecting this freedom in the context of AI "requires intervention from government and non-governmental organizations to ensure that individuals or minorities benefit from equal opportunities." However, the report does not discuss how AI systems can be used to support this goal, such as by reducing gender biases in recruitment processes. Moreover, it implies that companies employing AI systems are more likely to discriminate against certain groups. Chapter 1, Section 3.3 ("Respect for democracy, justice and the rule of law") asserts that AI systems destabilize democratic processes and societies, and "undermine the plurality of values and life choices." But again, these claims are not made on the basis of careful research and review of evidence. Moreover, such unfounded claims will not help nourish trust in AI from users, negatively impacting social acceptance of AI and, in turn, slow down the adoption of AI technologies. This section also would require AI systems to take on responsibilities that would be impractical. For example, the HLEG says AI system could "abide by mandatory laws and regulation, and provide for due process by design." Without mentioning which laws and regulations—and whether they are local, national, regional, or global ones—they refer to the "right to a human-centric appeal, review and/or scrutiny of decisions made by AI systems." However, the guidelines do not specify what this "right" would entail or how it could be operationalized, setting up a vague standard that most businesses will be unable to commit to. Chapter 1, Section 3.5 ("Citizens' rights") rejects all types of "systematic scoring by government," to which "citizens should never be subject." Many scoring systems have long been widely used throughout the EU, such as for credit ratings, and should not be dismissed out of hand. For example, most educational systems, including in EU member states, use scoring systems, and these scores may be biased by the judgment of a teacher. Yet AI systems could reduce the level of subjectivity in grading assessments and other types of scoring systems. To be sure, governments can abuse such systems, as the Chinese government is doing with its social credit scoring system. But that should not be used as an attack on the technology any more than steel technology should be

more arbitrary, and more susceptible to bias than algorithmic decisions—which is the reason why many organizations choose to adopt AI systems in the first place. Humans are also far more like "black boxes" than are algorithms, which heightens the folly of subjecting human decisions to lesser scrutiny than algorithmic decisions. In most cases these systems are less biased than human decision making, where subconscious or overt biases permeate every aspect of society. It is certainly true that AI systems, like any technology, can be used unethically or irresponsibly. And combating bias and protecting against harmful outcomes is of course important. But those who resist AI based on this concern fail to recognize a key point: AI systems are not independent from their developers or the organizations using them. If an organization wants to systematically discriminate against certain groups, it does not need AI to do so. A more constructive approach would be to recognize that human decision-making is subjected to less scrutiny than AI yet operates within "black boxes" of its own and greater use of AI could mitigate some human biases. Fourth, with respect to privacy, the HLEG fails to identify opportunities to use AI to increase individual privacy, such as by automating certain processes that would otherwise require an individual to reveal personal information to another individual. AI offers an important opportunity to increase privacy, and the HLEG should identify some of these opportunities where AI has a net positive impact on consumer privacy and encourage those uses.Fifth, the use of broad language and unclear terms is concerning. For example, the guidelines (see Chapter 2, Section 1.7, "Respect for Privacy") mention the importance of companies fully complying with the GDPR "as well as other applicable regulation dealing with privacy." The HLEG should specify which other applicable regulations the guidelines are referring to so as not to leave this open-ended to possibly include future regulations or ones in other countries. The HLEG states that adoption of these guidelines should be voluntary, but the guidelines recommend "formal" mechanisms, frameworks, constraints, procedures, and regulation. Moreover, the guidelines include references to "requirements" which could suggest there would be consequences to non-endorsement and non-adherence. Finally, the guidelines call for transparency and explainability, but make no distinction between the two. The two terms are commonly conflated in discussions about governing algorithms, and the guidelines reflect this particular misunderstanding as well, as they define "explainability—as a form of transparency." Transparency refers to disclosing an algorithm's code or data (or both), while explainability refers to the concept of making algorithms interpretable to end users, such as by having operators describe how algorithms work or by using algorithms capable of articulating the rationales for their decisions. The guidelines should clarify this distinction. Moreover, while transparency and explainability are fundamentally different concepts, they share many of the same flaws as a solution for regulating algorithms. In particular, they hold algorithmic decisions to a standard that simply does not exist for human decisions. If an evaluation of their decision-making

technologies or processes. AI is likely to be deeply integrated into organizations, and it will likely not be possible to always treat AI accountability and ethics questions separate from other organizational accountability and ethics questions. The draft oddly categorizes "ethical oath" as a skill and knowledge, according to another question listed under "Accountability." For the requirement "Data governance," the question "Who is ultimately responsible?" implies that organizations can easily and clearly determine who may be liable, which may not always be the case. It would be relevant to add some elaboration to this question, such as "Who is ultimately responsible for X part of process Z?" For the requirement "Respect for (& Enhancement of) Human Autonomy," the requirement for businesses to offer users the possibility to "interrogate algorithmic decisions in order to fully understand their purpose, provenance, the data relied on, etc." may be impractical for many organizations. In addition, the HLEG wrongly associates "risks to mental integrity" with "nudging." "Nudging" remains undefined, and could broadly include any recommendation, therefore including it in the list is not appropriate or practical for an assessment.

presents about AI is negative and unbalanced, especially given the vast number of tangible examples of AI's benefits already in existence and the relatively few instances of substantial and unmitigated AI harms from systems that have actually been deployed (as opposed to being tested). Indeed, there are several examples in the document suggesting that AI has greater potential to cause harm rather than to produce benefits. For example, in Chapter 1, Section 3.3 ("Respect for democracy, justice and the rule of law"), the HLEG suggests that AI systems "interfere with democratic processes" and "undermine the plurality of values and life choices." Such allegations are not supported by evidence and stand to diminish public acceptance of AI, which would slow down adoption. In contrast, Chapter 1, Section 3.5 states that AI systems only "hold potential" in terms of how they can "improve scale and efficiency of government in the provision of public goods and services to society." This statement mischaracterizes the numerous examples of AI systems already in production in governments around the world, while significantly overstating actual real-world AI harms that have occurred. There are two reasons why harms are likely to be vastly less than portrayed. The first is that in existing EU laws and regulations would apply to most applications of AI, giving governments the right to bring action against potentially harmful cases. The second is that those laws and regulations, along with oversight by civil society and pressures from market forces (e.g., the desire of companies to sell AI applications and maintain healthy public reputations) will lead the vast majority of companies to work diligently to ensure that the AI systems they deploy are accountable and beneficial. To address this shortcoming, the HLEG should provide more representative descriptions of AI's capabilities, and a clearer acknowledgement of where it is already delivering benefits and where concerns are merely speculative or have occurred but have been easily remedied. In particular, the HLEG should focus on informing the public about many of the positive use cases of AI, including industry-specific examples, as this will help create an environment that is more conducive to adoption of AI to the benefit of EU businesses, consumers, and others. For example, a negative tone could prove harmful to the development of a workforce with the technical skills that will be necessary for AI in Europe. European students will be unlikely to pursue a career in AI or related fields if those who contribute to its development are demonized. The HLEG's guidelines should not discourage policymakers from responding to legitimate concerns and discussing challenges, but they should also not encourage alarmists to delay progress. Second, for all of its concern about the future of AI, the HLEG ignores the fact that the EU is unlikely to be able to influence global AI ethics if Europe is not a leader in AI development and adoption. Ensuring "technological mastery" to foster "trustworthy AI"—an objective the draft guidelines set forth—requires the EU to be a global leader in AI. Europe is facing intense global competition in AI, but the HLEG ignores the need for the EU to focus on boosting public and private sector investment, raising technical skills of its

criticized because totalitarian regimes use steel to build prisons holding political prisoners. This section also suggests that AI systems only "hold potential" in terms of how they can "improve scale and efficiency of government in the provision of public goods and services to society." Yet there are many examples of how government are using AI systems effectively, and there is widespread agreement among AI experts that these systems will be even more impactful going forward. The introduction of section 4 includes vague language such as "in particular situations" or "Given the potential of unknown and unintended consequences of AI." This should be clarified. Chapter 1, Section 4 ("Ethical Principles in the Context of AI and Correlating Values") provides a number of potential ethical principles but does not elaborate on how organizations are already using AI for these goals. For example, the HLEG writes that "AI systems can be a force for collective good" but gives few details on this under its description of "The Principle of Beneficence: 'Do good.'" Other principles, such as "The Principle of Non Maleficence: 'Do no Harm'" which states that "AI systems should not harm human beings," are aspirational, but unrealistic. For example, if an organization truly abided by this principle to never cause harm, it could never use AI to eliminate a particular worker's job, even if on net workers came out ahead through higher living standards, or use AI in for autonomous vehicles that might result in human injury, even if on net there were many fewer accidents and injuries. Similarly, "The Principle of Autonomy: 'Preserve Human Agency,'" provides no explanation of how a "right to opt out and a right of withdrawal" can work in practice for certain uses of AI, such as facial recognition, where individuals may not have an interface to the technology. The draft guidelines are also vague about what it means to have "a right to decide to be subject to direct or indirect AI decision making," or what qualifies as an "indirect" decision. This also sets up a false comparison as there are a vast array of situations in Europe where individuals are subject to decisions where they do not know the reasons behind a decision (e.g., being accepted to a college, obtaining a job, getting a loan, etc.). Similarly, in "The Principle of Justice: 'Be Fair,'" the directive that data practices be aligned with "individual or collective preferences" is quite possibly unachievable, as there are as many preferences as there are individuals, and the collective preferences may not reflect individual ones. Likewise, this principle says that "the positives and negatives resulting from AI should be evenly distributed" which again may be aspirational, but not a standard that can be perfectly achieved and one that is not expected for human-led processes. Finally, "The Principle of Explicability: 'Operate Transparently,'" overemphasizes the importance of auditability and explainability, even those these requirements can limit the use of more accurate algorithms and undermine attempts to protect intellectual property by forcing companies to disclose source code. Having to explain the logic behind algorithmic decisions to as broad an audience of users as possible is an impractical requirement that could compel companies to make trade-offs between accuracy and interpretability of process happens at all, humans are rarely asked to explain is prior to the decision. In addition, mandating that companies make their propriety AI software publicly available would prevent companies from capitalizing on their intellectual property and future investment because other companies would simply copy their algorithms. Similarly, requiring explainability will limit the use of AI in Europe, and thus related investment, which will likely slow down research dedicated to this purpose. As a result, these guidelines could paradoxically act against their own advice by slowing research into AI.

workforce, and designing a regulatory environment conducive to AI so that it can compete with countries like China and the United States. For example, leading AI research is coming from North America and China where large tech companies have set up their own AI research labs because they have better access to talent, funding, and data. In addition, EU regulators have not been sufficiently supportive of AI. For example, regulators should foster voluntary data sharing to increase access to valuable data sets that may enable advances in machine learning. Often, the public and private sectors hold valuable data but lack mechanisms to securely and efficiently share it. Moreover, some provisions of the GDPR limit data collection and sharing and include other measures that will limit AI adoption. Amending the GDPR to ensure it does not impede innovation should be seen as a priority. Yet the draft guidelines rarely refer to the importance of increasing R&D, improving workforce training, or reforming regulations to make the EU more competitive in AI. The HLEG should draw attention to the fact that Europe is lagging in all three areas and should identify these priorities as a necessary precursor to influencing the global debate on AI ethics. In short, it is much easier for leaders to influence the overall direction of AI ethics, not only through market leadership but also through technological capability. Third, the HLEG's guidelines naively suggest that "user trust" will enable Europe to be globally competitive in AI. This, to be blunt, is wishful thinking that is not supported by evidence or real logic. Past studies that have quantified user trust in digital technologies have found that the levels of consumer trust in the EU are similar to those in the United States, even though the U.S. privacy regulatory system is not as stringent as Europe's. It is not that well-established that user trust—beyond a baseline level—deters digital adoption, and there is little evidence that user trust will be a major driver of AI adoption. What will be the major drivers of AI adoption will be the innovativeness, quality, cost, effectiveness and breadth of AI applications. Fourth, the HLEG incorrectly presents AI and ethics as a trade-off. For example, throughout the text, the draft guidelines suggest that an increased use of algorithms would lead to a host of harms, including exacerbating existing biases, discrimination, and inequalities. If the HELG is going to present such claims, it needs to thoroughly document them with more than assertions from civil society groups with an interest in limiting AI adoption. Moreover, it needs to examine all claims of harm not just from a first-order perspective (e.g., did a particular version of AI lead to troubling or problematic results), but from a second-order perspective as well (e.g., did the next version of the AI application fix that problem? did the application lose out in the marketplace to other applications that did not have that problem? etc.). A major problem with making these accusations, and implying that AI is inherently problematic, is that it will engender support for policies to regulate algorithms in ways that would harm consumers, businesses, and democratic values alike. Combating bias and protecting against harmful outcomes is important, but it should be made clear that if an algorithmic system produces unintended and potentially

their computer models. This section also fails to acknowledge the important research advances that might allow future AI systems to provide explanations. For example, the U.S. Defense Advanced Research Projects Agency (DARPA) is investing heavily in its Explainable AI program to spur breakthroughs in machine learning techniques that could explain themselves or be more interpretable by humans without sacrificing performance. Explainable AI would be enormously beneficial for applications ranging from judicial decision-making to medical diagnostic software, and would alleviate pervasive concerns about the potential for AI to be biased and unfairly discriminate. Rather than call for companies to use explainable AI before it has been fully developed, the HLEG should call for more research in this area and limit requirements for explainable AI to instances where accuracy is not more important. Chapter 1, Section 5 ("Critical concerns raised by AI") acknowledges that "our understanding of rules and principles evolves over time and may change in the future." This point is important, and since rules and principles are not timeless, the EU should be cautious about imposing static regulations on such an early and dynamic technology. Rules, principles, and concerns will likely change in the future, but regulations tend to lag behind technological developments. Therefore these guidelines should not mandate strict government standards. The HLEG should also clarify whether there may be any consequences for those organizations that do not choose to endorse these guidelines. It should also refrain from using language such as "requirements" given that this is intended to be a voluntary set of guidelines. Chapter 1, section 5.1 ("Identification without consent") refers to facial recognition as an example of "involuntary methods of identification using biometric data" and recommends overhauling the mechanisms through which consumers give consent, arguing that they are ineffective because "consumers give consent without consideration." First, facial recognition is not always involuntary, and so the guidelines should be updated to clarify this point. Second, it is not practical for consumers to give consent to many uses of facial recognition, such as when it is being used in a public place for public purposes, so that should not be the standard. Chapter 1, Section 5.2 ("Covert AI Systems") suggests that AI systems are necessarily risky and therefore people should have a right to know when they are interacting with them. As this requirement presupposes that AI systems pose some kind of inherent risk, it should be eliminated, and the guidelines should explicitly avoid rules that discriminate against the use of AI systems. Moreover, such a requirement will seem anachronistic in a decade or two when AI is used to improve significant parts of people's daily lives. Chapter 1, Section 5.4 ("Lethal Autonomous Weapon Systems (LAWS)") raises concerns over the "unknown number of countries and industries" which are actively "researching and developing" LAWS. Europe should begin to understand the potential military applications of AI. Rather than sitting back while other countries explore these uses of AI, Europe should work to understand its potential use by and against adversaries, especially to protect its

discriminatory outcomes, it is not because the technology or the developer is malicious. Rather, unforeseen limitations in the design of the system or reflections of real-world biases from training data may cause these types of errors, something one would expect with any new technology or system where developers are still learning and improving. But even where bias in AI systems may occur, in many cases, these systems are still likely to generate less bias than similar human processes. In addition, these biases can be identified and quickly improved, which is exactly what occurs in virtually all identified cases in the marketplace. Indeed, rather than treating AI as a technology that presents inherent ethical risks, the HLEG's draft guidelines should focus more on how AI could be used to address existing ethical problems by automating activities where humans have a propensity to act unethically, often unconsciously. Finally, the HLEG should eliminate some of the principles and requirements it proposes in the draft guidelines, such as transparency and explainability. By proposing these concepts as requirements for AI systems, they would hold algorithmic decisions to a standard that simply does not exist for human decisions and limit the use of some advanced algorithms that cannot easily be explained but offer greater accuracy. In addition, transparency requirements could entail code disclosure. The economic impact of asking for companies to reveal their source code would be significant as it would prevent them from capitalizing on their intellectual property and future investment, and AI R&D would slow because businesses could simply copy the work of others. A better alternative to transparency and explainability is algorithmic accountability—the principle that an algorithmic system should employ a variety of controls to ensure the operator can verify algorithms work in accordance with its intentions and identify and rectify harmful outcomes.The draft guidelines, while well-intentioned, miss the mark in terms of outlining a path forward for how the EU can be a global leader in AI, and through this leadership, answer important ethical questions about the future uses of AI. Rather than attempting to proceed on its own at setting global norms on AI ethics, the EU should work to establish itself as a leader in AI development and use, and work with other countries to develop common baseline approaches to AI ethics.

				infrastructure and strategic interests. But decisions about whether to pursue LAWS should not be part of the HLEG's mandate as it encompasses many broader questions about regional and national security that are outside the area of focus of the HLEG members. To that end, the guidelines should avoid conflating the broader debate about AI ethics with calls for banning "killer robots." That may be an important debate, but it is almost completely separate and distinct from the one about how AI will impact Europe's economy and society as a whole. Should policymakers succumb to baseless fears that military AI research will lead to a dystopian world full of rogue systems taking over the world, it will set back important AI research poised to deliver many benefits to Europeans. Debating how nations should govern and use autonomous weapons has its place in policymaking, but the HLEG should be careful to recognize that this technology is not just about "killer robots." By comparison, policymakers in the early 20th century did not conflate debates about the internal combustion engine with questions about using that technology to power military tanks. Sabotaging important AI research that can serve the public good as a means of avoiding confronting these issues head on is counterproductive and will harm innovation.The HLEG states that it will add a final section (5.5) to explore "Potential longer-term concerns." The draft guidelines note that this section is "highly controversial" within the HLEG itself. Given that these concerns are so speculative as to be closer to science fiction than science, such as positing risks from AGI, they should be excluded from this report. As noted by AI expert Max Versace, CEO of robotics and computing company Neurala and founding director of the Boston University Neuromorphics Lab, "The likelihood of an AI scientist building Skynet is the same as someone accidentally building the space station from Legos." If the HLEG decides to include these purely speculative long-term risks, then it should also include a similar section outlining the potential long-term benefits of unforeseeable advances in AI.		
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	
Claudia	Otto	COT Legal						Please see https://cot.legal/cot-legal-ai-statement-en.html

Jelle	Hoedemaekers	Agoria	Agoria is the Belgian federation for the technology industry. We are paving the way for all technology-inspired companies in Belgium pursuing progress internationally through the development or application of innovations and which, together, represent some 300,000 employees. We are proud that more than 1,900 member companies place their trust in the three pillars of our services: consulting, business development and the creation of an optimal business environment.Agoria supports the initiative taken by the High Level Expert Group on Artificial Intelligence in writing Ethical guidelines for AI. We believe that it is of great importance to promote the development and adoption of a European competitive and trustworthy AI, and therefore welcome the opportunity to submit our comments. Glossary"Artificial Intelligence" : The definition is very detailed, but it indicates an application field more narrow than other existing definitions, too narrow in our opinion. Specifically, there is too much emphasis on the aspect of decision-making ("decide the best action to take") and it is unclear what this would mean in practice. Is the scope limited to systems that make decisions, excluding systems which leave this aspect to a human being (either partially or entirely), enabling it by generating certain findings or even suggesting options for resolving problems? If the decision making is seen as an important element of the definition, then clarification is needed on what a decision like this would be, and what would constitute "the best action". Finally, in the GDPR, article 22 references "decision based solely on automated processing". In our view, the High Level Expert Group should consider the relation between this definition and article 22 in the GDPR."Bias" : The current definition of "bias" appears too narrow, given that bias is something that will always be present in datasets, depending on the position from which it is viewed. Additionally, the existence of (some) bias should not always be considered a bad thing, but sometimes it is intended in the use of the AI. What is to be avoided is discrimination based on a biased dataset, and according to us this is is important and well reflected in the ethical guidelines. However, we think it would be beneficial to view bias more from a technical point of view and distinguish it clearly from unfair bias. The goal should not be eliminate all bias in datasets but to help people understand the scope and limitations of the dataset and to take steps to mitigate the risk that an AI solution will generate and apply an unfair bias.Rationale and Foresight of the GuidelinesWe support the Trustworthy AI approach and the fact that it would consist of the condition of an "ethical purpose" and the requirement of AI being "technically robust". In light of the need to foster innovation, we also appreciate that the High Level Expert Group does not propose a drastic regulatory approach, but instead proposes overall principles that may need to be finetuned over time. The mechanism to enable stakeholders to endorse these principles is highlighted. We understand that this is the ultimate goal, and we may consider motivating our members to endorse these guidelines. However, in its current form the document is too vague for it to generate sufficient interest in the Belgian industry to endorse it or sign up. It needs to be clearer	We broadly agree with the principles laid out in this chapter. The fundamental rights approach is a good way to define the principles and values. We would advise that this section is adapted to include more guidance towards evaluating AI, and make it clear that sometimes an analysis of AI from a fundamental rights, principles and value perspective will be more of a balancing exercise of the benefits and harm of AI systems and not only a general indication of the implications of the use of AI on existing fundamental rights. Additionally, the question arises of what would happen if there would be a contradiction between principles which are derived from the fundamental rights? Is there a hierarchy which should be followed in this case? Guidance in this matter would be required.We believe that the document focuses very much on AI based on personal data, hence on issues such as informed consent and opt-out. It is however important to note that, especially in many B2B applications, AI may not be based on personal data. In that sense, the document is in our view over-emphasizing the relation between AI and personal data. Also, the question of legitimate processing under the GDPR is still different from the questions around data quality, possible bias and trustworthiness. And finally, while we do not contest that in some cases informed consent will be the adequate legal ground for lawful processing of personal data, it cannot be excluded that in other cases processing of personal data can be validly done on the basis of 'legitimate interests', 'further compatible processing', 'vital interest' or the other legal grounds foreseen in the GDPR. We think that the ethical and human rights assessment of AI systems should take into account and build upon the privacy impact assessment required under GDPR rather than requiring an additional privacy assessment exercise. From that perspective, we also recommend that the focus in the guidelines should be more on creating awareness about how personal data are used in an AI context, increasing awareness about the impact of AI on privacy while also providing the reassurance that AI systems that follow the guidelines are trustworthy and beneficial to individuals and to society. An initial risk assessment could help to clarify if an opt-out is a heavy requirement and therefore required in the AI system.The operate transparency principle should reflect more that this is about informing individuals about whether or not they are interacting with an AI system. We are unsure about the practical realization of 'comprehensible and intelligible by human beings at varying levels of comprehension and expertise' AI systems. Does the High Level Expert Group feel that this would be a prerequisite for all Trustworthy AI applications? Relating to the critical concerns section, specifically regarding identification without consent, we recommend that the guidelines expressly acknowledge that different applications of AI might warrant different types of consent. This should be linked to the potential individual or societal harm.	This chapter gives a good start on a guidance for implementation, but the 10 requirements mentioned make it quite extensive. We would prefer to see less requirements, for example by combining some of them.For example : "data governance" and "respecting privacy", "design for all" and "non-discrimination", "governance of AI autonomy" and "respect for human autonomy" , "robustness" and "safety" could be combined in our opinion.The 'design for all' principle raises certain questions and may set the bar too high. For example, if there is not enough data available on a specific subgroup, the decision made by the AI cannot be motivated for said subgroup due insufficient training data. Would it not be better to exclude this subgroup rather than making a faulty, unsubstantiated judgment? In this case we assume that the system would be accessible for all, but not applicable for all. We would recommend to clarify this and add this nuance to this section. We have doubts whether the 'design for all' principle can be implemented in practice.We would prefer the 'design for all' principle to be defined as suggested by DigitalEurope as ; "Systems should be designed in a way that considers usability and accessibility so that the products or services should be inclusive and can be accepted by as many citizens as possible, regardless of their age, disability status or social status". In the robustness section we would like to see guidance on required accuracy for AI systems. Is there a plan to set up a guidance process by the High Level Expert Group or the European Commission? The fallback plan section could be more detailed. According to us this should depend on the use cases and in some situations it might not be required. In the transparency section, the "development processes" should be clarified more. It seems to refer to system design processes, and it would be unrealistic that companies would be fully transparent in their design process as this is what gives them their competitive edge.In our view, the level of transparency of an AI system is something that is use case specific and hence should be considered in relation to the different types of use cases.We support the decision to list technical and non-technical methods for achieving trustworthy AI and recognize that actions are being undertaken to develop these methods.	This is a good first step but we would like to have a more clear and detailed list that can practically be used by companies involved in developing AI. In any event, we consider that the list of principles should be use case specific.In the final guidelines we would like to have a more clear view on technical means to assess when an AI product would be fit for launch, or how to judge whether it would be fit for launch or not. We understand that this requires substantial technical input, but if there would be a possibility for endorsing this document, systems for assessment and confirmation need to be set up.	In general we agree with several of the proposals contained in this document, however we feel that the focus is too much on "AI for consumers" and the processing of personal data. We think that the High Level Expert Group should consider more Trustworthy AI in relation to its purpose and intended users. This implies a risk-based approach during the early development in which the developer considers the potential impact or harm in reference to the intended use and users of the AI system. Based on this assessment the required level of ethical purpose and technical robustness can be set. If this mechanism would not be included we fear the this might hinder the development and adoption of AI in an industrial or B2B setting. If such a mechanism would be included, this would help the developers relying on these guidelines to apply them appropriately. This would also imply that in some cases, the assignment of an ethical expert may be inappropriate and irrelevant.Additionally, the process of endorsing these guidelines should be clarified. It is often stated that these are guidelines and non-mandatory, but the document often mentions words such as "compliance" and "requirements". We would like to note that in the final version where the voluntary endorsement mechanism will be specified, the legal consequences of this endorsement should be carefully considered. For example, the endorsement could have consequences taking into account the existing EU directive 2005/29/EC on unfair business-to-consumer commercial practices, in particular Article 6.We appreciate the work that has been undertaken by the High Level Expert Group in establishing these guidelines for Trustworthy AI and look forward to engaging in further discussions.

						to industry players what they would endorse or sign up for: what are the implications for possible signatories? What would be the required follow-up in the future? What would be the legal consequences of an endorsement, if any? We support the assessment made by the High Level Expert Group in the Scope of the Guidelines that there already are requirements to comply with fundamental rights and applicable regulation which are or could be applicable to AI. It is our view that currently there is no legal vacuum, nor requirement for specific general AI regulation.We agree that all stakeholders need to be informed and ought to be involved if we would want to move towards Trustworthy AI.We strongly agree that a tailored approach for development of Trustworthy AI is required based on the context of its use cases.
Meeri	Haataja	On behalf of open workshop organized by Finland's AI Program's Ethics Working Group				On behalf of an open workshop organized by Finland's AI Program's Ethics working group, we're raising the following notes on the HLEG Draft Guidelines for Trustworthy AI: 1. Appreciation for the strong strategy of EU: EU can and should take an active and strong strategy on the ways we are building our society in the AI era and towards long-term impacts of AI in our society. 2. Recognizing the existing laws and regulations: Report could make the existing regulative frameworks clearer as they provide the definite ground for ethics implementation. The recommendations should guide towards GDPR-compliance. 3. Data ethics and ethical data: More emphasis on how data to develop AI is captured and guidelines on reuse of data could be added to the paper. Also, the data quality has an impact: poor data quality can result in unethical AI. 4. Concept of MyData: In order to enable availability of ethical data, MyData should be recognized as an enabler for ethical and human-centric AI. Clear guidelines on enforcing MyData in align with GDPR would be highly valuable. 5. The power of the developers: The question on the power of the developers rises especially from the startup environment and culture: A great impact and power can be reached even with really limited developer/tech groups and teams. 6. Transparency: Demand for clarity on the dilemma of trade secrets vs. transparency, from the point of view of competitiveness of European companies. Transparency should not be regarded as a requirement for open sourcing all AI. 7. Design for all. There was discussion on the business-centricity of the paper and that the “design for all” recommendations should consult more perspectives from e.g. NGOs. 8. Inclusivity and design: Inclusivity is a recognized principle for all AI development. Group raised the report could offer a principle-level—practical case examples and guidelines. These examples could introduce the readers to the existing guidelines and

							principles of inclusivity in its every form. Also here, guidelines would value of consultation with NGOs. 9. From principles to guidelines and practice. Report claims to provide guidelines, but in practice provides principles on methods to achieve trustworthy AI. These principles are considered good, but in addition there is a need for more practical guidelines for implementation with relevant examples. Also, a need for different versions for different target groups was raised (education, politics, business, non-expert). 10. Special considerations related to reinforcement learning. While supervised learning seemed well covered in the paper, there was some concern that it did not sufficiently address challenges related to reinforcement learning. Due to the special nature of reinforcement learning, questions on e.g. auditability and transparency need more attention and should be addressed separately. On behalf of an open workshop organized by Finland's AI Program's Ethics Working Group, Meeri Haataja Ethics WG Chair, Finland's AI Program
Jenni	Hyppölä	CSC – IT Center for Science	The objectives and goals set for the ethical and sustainable development and application of AI in society are well-contemplated, just and honourable. In the wider global view, it is also prudent to strive towards the branding of human-centric "Trustworthy AI made in Europe", as opposed to the mainly consumer-centric American AI or the government control focused Chinese AI. AI pursuits and developments are already embodied in a large number of everyday contexts. Now is the last moment to foster true reflection and discussion on an ethical framework for AI. This should be done also at the global level. Europe should be a driving force in discussing the ethics in AI internationally. Global rules, based on ethical values, should be established, and EU should be an initiator in this discussion. Principles and good practices adapted in the forthcoming final version of the "Ethic Guidelines for Trustworthy AI" will be a good starting point for the discussion on global ethic guidelines for AI. The integrating European Union has also previously set goals and expectations on the role of Europe at the global scale or level, the realisation of which has sometimes fallen short in implementation and been hampered by various nationalist policies. Thus, Europe has been in a disadvantaged position in the global competition. AI is developed everywhere with a knowledge-based set of technologies par excellence. It is thus imperative to coordinate joint European-level AI development efforts and not to be misled by assuming the other global players would not already have a technological edge and lead in Ai. Notwithstanding, the ethical point of view of core principles and values is still precisely the correct and the best bet that Europe can bring to the table of discussion on AI at the global level.	AI, as most technologies as such, is value neutral. It can be used for either good or bad purposes. The usage and its ethicality depends on human beings developing and using AI. The fundamental rights described in the Chapter I are the basis for the Draft Document. This approach is highly favoured. Until now, the development of AI has been dominated by technological and commercial interests. Ethical concerns have been addressed once problems have arisen. Legislation and regulation always lag behind technical development. The Draft Document correctly points out we require "…guidance on what we should do with the technology for the common good rather what we (currently) can do with the technology". On the other hand, legislation must be renewed and updated so that it also enables the full potential of AI for good purposes. E.g. in the upcoming Copyright Package data mining should be enabled to wider extent than now is proposed. The ethical principles and correlating values presented in Section 4 are to be favoured. However, one should bear in mind that sometimes it might be difficult to determine, how to define e.g. "good" in a certain context. If someone's "good" is less "good" for someone else, whose good will be more respected? Values may also be in conflict with each other sometimes. This is often the case e.g. when creating common security requires limitation on individual freedom. The AI HLEG asked for specific input on the Section 5. It should be noted that AI is already used for purposes that are in conflict with the fundamental rights or can be seen unethical. Identification technologies are already being used for identifying people and there are hidden attempts to affect on people's opinions and democratic elections with AI systems. Using AI for scoring citizens and societal control system is reality. Europe is not safeguarded from this kind of attempts, nor are the Europeans. A possibility of opting out is mentioned in the	It is important to increase the education on data science in all levels of education. The expertise is yet not enough. In the modern world, all citizens must be given prompt training on digital skills as a part of the common knowledge. Everybody should also be taught to understand the value of their personal data. Data skills should be also included in the key skills of lifelong learning. It should be noted, that AI is based on code and algorithms written by people. Thus, data science studies should include ethics so that future experts will have understanding on ethic related issues in data handling and use of data.Non-discrimination is an important principle. It is important to understand that algorithms may cause unintentional harm in the real life, if potential risks or side effects are not recognised while creating algorithms. A case example of this might be an AI assisted system that ranks job applicants for an interview based on their applications. Unintentional discrimination may also remain unnoticed. Thus, it is important to carefully plan and simulate algorithms before taking them in use especially in public services.	Based on common values, the EU is a natural actor to promote ethical use of AI. The assessment list presented in the Draft is a good reference point when assessing Trustworthy AI. As the document states, the list is not exhaustive and assessment is a continuing process. At this stage, it remains unclear, if Trustworthy AI should be self-evaluated by an organisation, or if an external auditing should take place. In both cases, Trustworthy AI should have valuable and wanted status. There are many examples of national, regional, or international rating or certification systems that provide prestige to both organisations and consumers, such as Fair Trade, FSC (Forest Stewardship Council) for sustainable forestry), and MSC (Marine Stewardship Council for sustainable fishing) certificates. Should there be a suchlike organisation to evaluate and credit organisations following the Trustworthy AI ethic guidelines? Such a system might promote ethical use of AI and encourage organisations, perhaps also other global actors, to make their systems more compatible with the Trustworthy AI principles. This has been experienced e.g. with the Bologna process in higher education, as non-EU countries have started reforms to adapt their systems to be more compatible with the European education system.	CSC – IT Center for Science is a Finnish center of expertise in information technology owned by the Finnish state and higher education institutions. CSC provides internationally high-quality ICT expert services for higher education institutions, research institutes, culture, public administration and enterprises to help them thrive and benefit society at large. CSC supports the EC's HLEG AI work to compile ethic guidelines for trustworthy AI and thanks for the opportunity to comment on the draft version and working document of the guidelines.

				document, but it is not clear how this opt out would be made possible in large data sets collected by different public and private actors. However, the Europe should actively promote everyone's control over their own personal data, based on MyData approach.In terms of the longer-term concerns, it is highly likely that radical technological changes will take place in the future. For many concerns and risks presented in the document, the question is not whether but when will they realise. Legislation and conventions always lag behind technological development so it is wise to address all known concerns as they appear and be prepared also for unpleasant and unlikely scenarios. The history has shown that if something is possible, it will be used, unless it is regulated by international conventions. Examples of such successful conventions include e.g. prohibition of chemical and nuclear weapons, which have mainly been widely accepted and effective.			
Alejandro	Moledo	European Disability Forum	Page 4. In the Executive summary, Artificial Intelligence (AI) based solutions improving accessibility for persons with disabilities can also be highlighted, since access to Information and Communication Technologies is definitely a "grand challenge" for 15% of the population (80 million Europeans with disabilities). However, the assertion that "Given that, on the whole, AI's benefits outweigh its risks" is something we cannot state at this stage. Users can find applications based on AI which improve their lives, but others may also discriminate against them (e.g. AI-based recruitment procedures, insurance price setting). Glossary. The definition of Human-centric AI approach must recognise human diversity: "The human-centric approach to AI strives to ensure that human values and diversity are always the primary consideration, and forces us to keep in mind that the development and use of AI should not be seen as a means in itself, but with the goal of increasing all citizen's well-being.".	Page 14. Under paragraph "3. Fundamental Rights of Human Beings", on 3.4 Equality and non-discrimination, a clear reference to the United Nations Convention on the Rights of Persons with Disabilities (UNCRPD) is required. The UNCRPD was ratified by the EU and all its Member States and is binding on all state parties. Accessibility is considered a precondition necessary to enjoy the other rights enshrined in the Convention. As for the text of this paragraph, it would be best to address Equality as inclusion of "all people", and not just what societies may consider "minorities". This term may also be inappropriate when referring to consumers or workers. "Equality also requires adequate respect of inclusion of all people, including those traditionally excluded, especially workers and consumers." Page 16. Principle of Non maleficence: "Do no Harm", the following paragraph must be rephrased using respectful language: "People at risk of exclusion (e.g. children, [deleted:minorities], persons with disabilities, older people, or migrants) should receive greater attention to the prevention of harm and discrimination, given their characteristics and abilities. Inclusion and diversity are core aspects for the prevention of harm to ensure suitability of these systems across cultures, genders, ages, life choices, etc. Therefore, not only should AI be designed bearing in mind the potential impact on a wide range of people, but the above-mentioned groups should be taken into consideration in the design process (rather than just through testing, validating, or other)." Page 17. Finally, as for the Principle of Explicability, it must be added that "easy to understand information" is key for users to give informed consent. Over-complicated, liability-oriented texts do not help end-users, including end-users with intellectual disabilities. Page 18. Equally, the sentence "As current mechanisms for giving informed consent in	Page 21. The requirement Design for All must be followed by "and accessibility". Design for All is the European term for Universal Design. This must be clarified in a footnote. According to the UN Convention on the Rights of Persons with Disabilities, Universal Design means "the design of products, environments, programmes and services to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design. "Universal design" shall not exclude assistive devices for particular groups of persons with disabilities where this is needed". According to the General Comment number 2 of the CRPD Committee on the Convention article 9 on accessibility : "all new objects, infrastructure, facilities, goods, products and services have to be designed in a way that makes them fully accessible for persons with disabilities, in accordance with the principles of universal design" .It is important to understand that the right to access as described in the UN CRPD covers the full range of human diversity to any place or service intended for use by the general public. Denial of access is therefore an act of discrimination. Giving the aspirational nature of Universal Design applicable to everything new, it is essential that the document complements the requirement of Design for All with "accessibility". There are EU and national legislation setting out specific accessibility requirements that should be respected when developing a user interface or specific content (e.g. 2016 Web Accessibility Directive, and recently agreed European Accessibility Act). These legislations are underpinned by specific accessibility standards (e.g. EN 301 549) that must be followed when developing websites, applications or any kind of software. Even though Design for All principles should always be kept in mind, and there will be a European Standard on achieving accessibility following a Design for All approach (EN 17161), a specific mention of accessibility will give more certainty to the users of these Guidelines. Page 22: "3. Design for all and accessibility	Page 32. "3. Design for all and accessibility: Is the system equitable in use? Is the system flexible in use? Is it simple and intuitive to use the system? Is the information on the system perceivable, including for users of assistive technologies? Does the system arrange elements to minimize hazards and errors? Does the system allow for perceived errors to be corrected with ease? Does the system accommodate a wide range of individual preferences and abilities, including persons with disabilities? Does the system user interface follow the relevant accessibility requirements and standards and how is it verified? What definition(s) of fairness is (are) applicable in the context of the system being developed and/or deployed? For each measure of fairness applicable, how is it measured and assured? Were persons with disabilities involved in the conceptualisation, development, testing, implementation and monitoring as regards to the AI system?" Page 32. On questions about Non-discrimination, add: "Are measures in place to address cases of discrimination by the AI system?" Page 34. On questions about Transparency, add a question on understandability: "is the information provided to the user easy to understand?".	Artificial Intelligence will have huge implications on the lives of all citizens, including persons with disabilities who can already benefit of a number of AI-based applications. We nevertheless need strong legal safeguards to protect the rights of all citizens, including citizens with disabilities, from AI-powered technology that could cause them harm. In this context, an assessment of the potential gaps in human rights law that currently protect European citizens is required. Industry-driven guidelines on AI ethics are not sufficient to future-proof the rights of persons with disabilities. Self-regulation and voluntary compliance with ethics guidelines are not enough to offer reassurance to consumers with disabilities. We need a clear legal framework in place, clear accountability and a right to redress. The distinction between ethics and law must be clarified in the draft guidelines. We feel there is a risk of downgrading fundamental rights to ethics without any legal accountability. For example, who is responsible when AI is used in public decision-making? Linked to that, the document misses a crucial aspect: public procurement. The guidelines fail to address those who procure services and applications using AI, as often they are those who have more control and manoeuvre, compared to others (e.g. the end-users) in the process. These guidelines should also address the ethics of those procuring AI solutions Designers or those providing specifications are also hardly mentioned, including in the main diagram. These should also be addressed by the document, as they have the responsibility of key aspects such as usability, accessibility and the obligation to prevent discrimination. We also regret that the AI High Level Expert Group is only offering engagement opportunities on the European Commission Futurium "AI alliance" online platform. As highlighted by a member of our EDF ICT expert group, this online platform is not accessible for people who use assistive technologies. This means that the views of those citizens are not reflected in the online debates. This is discrimination by design. We are therefore urging the European

		the internet show, consumers give consent without consideration" is misleading, as it seems to put the blame on the user, when most of the current mechanisms to give consent are presented in a way that: first, is too complicated, and second, does not give choices to the user.		Systems should be designed in a way that allows all citizens to use the products or services, regardless of their age, gender, disabilities or characteristics. It is particularly important to consider accessibility of AI products and services for persons with disabilities as they are present in all societal groups. AI applications should hence not have a 'one-size-fits-all' approach and should consider Universal Design principles addressing the widest possible range of users, and should follow relevant accessibility standards (e.g. EN 301 549). Design for all and accessibility will be beneficial to all users, improving the usability of technologies for everyone anywhere and anytime, ensuring the inclusion of persons with disabilities in any living context, thus enabling equitable access and active participation of potentially all people in existing and emerging computer-mediated human activities. This requirement links to the United Nations Convention on the Rights of Persons with Disabilities" Page 24. On the Principle of Respect for Privacy, add "disability" as one of the sensitive data that users may want to keep private . Page 26. On the Technical methods - Ethics & Rule of law by design, add "Accessible by design", due to the fact that is required by the legal framework of the EU (UN CRPD), and because the European Accessibility Act will soon enter into force, and ICT companies will need to respect this legislation. Page 29. As for the non-technical, Stakeholders and social dialogue, civil society organisations should be added as one of the key stakeholders to set up the dialogue. Besides, on Diversity and inclusive design teams, "disability" must be included as one of the factors that should be incorporated in a diverse team. Finally, another non-technical method that must be added is Involvement of end-users throughout the whole design process; this would ensure that AI solutions do actually take into consideration the needs, requirements and expectations of all potential users.		Commission to ensure compliance with current web accessibility standards as a matter of urgency, as recommended by the European Ombudsman . Last, but not least, we would like to request the High Level Group on AI to follow its own requirement on Design for All and to ensure that the document is designed for all. The diagrams are not accessible to those using screen readers. Final note: an edited document is published on the Library of EDF website (www.edf-feph.org) including all the footnotes.

Asanga	Ranasinghe	STAMPEDE @SDGsTech Accelerator	The preparation for socio-economic changes under Pillar 2 should not be done only with a view of Europe. What about the rest of the world? Changes in Europe will influence changes all over the world, and in return, have a return-effect on Europe.How're the three pillars linked? For example, how will it be ensured that the investments under Pillar 1, especially private investments, are within the ethical and legal framework in Pillar 3?How will the so-called benefits of AI offset the environmental footprint of producing and using AI? Technology has a huge impact on the environment. The mining for minerals and raw material required to produce digital equipment takes a huge toll on the environment. When the consumer demand for technology is artificially created through greed-driven commercial principles rather than honest human principles, it depletes these natural deposits at an alarming rate. Further, a very large amount of energy is consumed for these mining operations to extract material. In addition, the processing of the material and production operations to create technology, not only consumes massive amounts of energy but also, creates harmful conditions for workers. It is inexpensive to produce technological devices in certain countries due to the lack of standards and implementation of labor laws. The factory workers have to work in conditions, which are harmful for their health due to the chemicals with which they come into contact. Long tedious hours and in certain cases, even child labor, are some of the despicable means employed by tech companies to make fat profits. This could even lead to cases of modern day slavery. The plastic packaging and energy-dependent logistics for distribution too has a toll on the planet. Finally, the current use of technology consumes and unprecedented amount of energy, which completely disregards people's right to safe and clean environments. While an estimated 1.1 billion people – 14% of the global population – did not have access to electricity1 according to Energy Access Outlook 2017, the elite consume more energy than necessary due to the use of technology. The situation has also affected other life species on land, as well as life under water, by destroying and diminishing their habitats. UN Digital Cooperation should make sure that principles, which protect inter alia SDGs 12, 13, 14 and 15 are considered seriously when addressing digital issues.How will ethics be used to inspire trustworthy development of AI in other countries, such as China, to cater to the huge demand that can be expected to be created in Europe and the Occident? Trustworthy AI Made in Europe must be comprehensive; it must ensure that processes and systems elsewhere in the world to develop, deploy and use AI in Europe follow the ethical purpose and technical robustness.	The interdependence of human life, other beings and phenomenon, within the earth sphere and in the universe, is subtle and unfathomable, especially to the distracted human mind. For example, human organ donation will reduce if AI cars create less accidents. This means lives are saved and unsaved at the same time (i.e. less accidents will save lives but, other human lives that need organs to be saved might not be so lucky). Who can decide what is right in this instance? It can also be argued that AI, along with 3D printing and other complementary technologies, will present solutions to those who need organs. The only certainty is that human lives will be deeply impacted by AI.General use of AI in social media has created distraction in society! Social media has affected people's productivity. Simon Sinek explains this very well: https://bit.ly/2Uuaok4When going through material about AI, at times, it feels as if though the global push for advancing AI is all about the money: the potential of $15.7trillion to the global economy by 2030 from AI (https://pwc.to/2hmUvOB). The EC must ensure that this doesn't contradict the declaration that the EU is upholding the human-centric values. Therefore, metrics to measure the human-centric impact from AI, such as expansion of choices available to people living in poverty, benefits to education, health, preventing domestic violence, crime etc. should be developed.Related to point 3.4 on page 7, Rights of different categories, such as Child rights, Rights of Indigenous People etc., should be considered, in addition to what is covered under the umbrella of fundamental rights. Therefore, relevant conventions, for example the CRC (Convention of Rights of the Child), Convention on the Rights of Persons with Disabilities (CRPD) and the United Nations Declaration on the Rights of Indigenous Peoples, should be considered as complementary documents to the Charter of Fundamental Rights of the EU and Treaty on EU.The human minds can predict and foresee the future only to a limited future, partly due to their greed. What about when AI demands Ethics? Should the inherent value of humans mentioned on page 5 be adapted to AI in humanoid form? i.e. Robots do not need to look a certain way etc. Will it be the same as for humans or different? Artificial Consciousness (AC), which are AI systems that may have a subjective experience, are a real threat to human autonomy. From a philosophical perspective, this might be the inevitable future of the human evolutionary trajectory. AC research labs in France, USA and Japan should immediately be shut down. Regarding Covert AI systems on page 11/12, an important question to ponder, even though it might not be in the immediate future, is if AI will develop human characteristics & face discrimination based on ethnicity, nationality, sexual preference, ability, age, gender etc.New trend in AI, Tabula Rasa, which is learning without data and human guidance, could undermine human autonomy and threaten human life. Intel's Ambient World envisions a future where the physical and cyber worlds converge. How are the masse affected by the choices of a few? Do citizens all over the world, 7 billion of them, demand this sort of a future? Are they even aware of these developments going on behind secret doors? Obviously not. How can	It's only matter of time before AI itself is used to test and validate intelligent systems. How robust would this be?In terms of XAI research, Hannah Arendt's thoughts on her 1958 work the Human Condition (HC) are useful to consider. As Arendt puts it: "The reason why we are never able to foretell with certainty the outcome and end of any action is simply that action has no end" (HC, 233). This is because action "though it may proceed from nowhere, so to speak, acts into a medium where every action becomes a chain reaction and where every process is the cause of new processes … the smallest act in the most limited circumstances bears the seed of the same boundlessness, because one deed, and sometimes one word, suffices to change every constellation" (HC, 190)." Non-technical methods of achieving trustworthy AI is important. But, we need to produce a cadre of people who have knowledge, both, of the technical and non-technical methods to achieve trustworthy AI.Consulting the work of the IEEE Standards Association might be worthwhile. "IEEE Standards Association (IEEE-SA) is a leading consensus building organization that nurtures, develops and advances global technologies, through IEEE. We bring together a broad range of individuals and organizations from a wide range of technical and geographic points of origin to facilitate standards development and standards related collaboration. With collaborative thought leaders in more than 160 countries, we promote innovation, enable the creation and expansion of international markets and help protect health and public safety. Collectively, our work drives the functionality, capabilities and interoperability of a wide range of products and services that transform the way people live, work, and communicate." Including the HDCA (Human Development and Capabilities Association) for stakeholder and social dialogue is recommended.There's a fake belief that we understand technology. Just because we use technology, we feel that we know it. For example, the internet. We use it but, do lay people really know what it is and how it functions? Do they know the difference between the WWW and the internet? Most likely not. So, people feel safe using it. And this goes for any other technology that might be harmful as well. People seem to feel a certain superiority in using technology. And unconscious reassurance that technology is your friend. This is a great advantage for technology producers.There are two separate issues that are connected. Firstly, nobody is saying that technology is evil. People are worried about the humans behind the technology, who might be evil; the 'evil humans' have the ways and means to manipulate the world to make profit and create inequality. Secondly, as technology advances, people are worried that AI will develop human like tendencies, which includes both good and evil. There already are examples where AI has shown to be aggressively competitive . Gmail's canned replies invert machine learning, so that automated replies "train the users, who function less as creative human beings & more as...neural nets that sift through AI-generated proposals & reject those that fail to conform to some pattern." This means technology too can become evil by imitating human characteristics, making the first fear of people void. This situation is worse and	While there are planning and operational tools to practice the DNH approach, the 'Four divine abodes' from the Buddhism offer a great set of philosophical guidelines. Buddhist texts translate the term brahmaviharas as "divine abodes," and state the four basic ones: metta (loving kindness), karuna (compassion), mudita (empathic joy), and upekka (equanimity). These four are attitudes towards other beings. They are also favorable relationships. They can also be extended towards an immeasurable scope of beings and so are called immeasurable . Can these be embedded in AI and used for its assessment?Tech companies are creating a generation of mindless zombies, who are glued to their digital devices; this needs to be prevented. Tech companies consider humans as commodities to further their profits. Instead, they should respect common principles. In fact, all stakeholders involved in AI should endorse the Principles for Digital Development and support its implementation.	The advice when it comes to developing, deploying and using is "Don't play God." We are not owners of the earth, its resources and, most importantly it inhabitants; human, animal and plant; on land, air and underwater. Everyone must act responsibly and respectfully as guardians and not owners of these to enable future generations to live peaceful and dignified lives. From an economic perspective (not purely), the market was decided by supply and demand of human needs. How will AI impact this?From a social and evolutionary perspective, autonomy of humans: Survival of the fittest was on a human scale. How does AI impact this?From a philosophical perspective, does the philosophical notion of Karma affect AI and robots? i.e. every action has a consequence.Can AI be prosecuted for a violation of conduct involving children? Ex: child abuse, show of porn etc. What about rape?World/Society is sending mixed messages: some are talking about ethics, principle etc. and protecting the rights of the robots; humanising robots. But, in VR, you can easily kill people and robots. References:https://stanford.io/2BfMhyChttps://bit.ly/2GltlSa https://hd-ca.org/ https://www.sciencealert.com/google-deep-mind-has-learned-to-become-highly-aggressive-in-stressful-situations https://tinyletter.com/robhorning/letters/reasons-to-believehttp://www.buddhanet.net/mettab5.htmhttps://digitalprinciples.org/

the EC influence these processes to be in line with the ethical purpose and technical robustness it advocates for AI? For Informed Consent to be effective, people need basic skills and knowledge, not just limited to literacy. In addition, people will need at least a basic idea and fundamental knowledge of technology; technical literacy. Young people and the generations born into an era of advanced technology might navigate with relative ease, if we envision an inclusive future, where everyone has access to technology. But, what about elderly people during the transition period to more technically advanced socieites? We are in that transition period right now.Long and cryptic legal agreements and user terms might be necessary to cover the technology producers but, this too undermines human autonomy, since consumers will not have the time and knowledge to read word for word and agree. Instead they will simply agree to the terms, driven by their eagerness to be served, blind enthusiasm to try the new technology and trust in the manufacturer/service provider. Richard Thaler's Nobel Prize winning research has shown that people can be cajoled to stay in a system by already including them and giving them the choice to 'opt out'. They will be reluctant to opt out.Human egos are fragile. What happens the moment someone who has the capacity to manipulate technology feels they were threatened, embarrassed or taken advantage of by someone else? Human or AI. It can safely be assumed that the person will use his tech knowledge to get back at the person or system who made him/her feel like that. This is applicable to world leaders as well. Therefore, future wars can be triggered and executed easily with unfathomable consequences because more than one party will have highly advanced digital arsenals. LAWS.Human dignity is of paramount importance. In response to a question raised by the former CFO of Yahoo at a Panel Discussion in Davos recently, the Executive Director of OXFAM mentioned how some women in poultry factories in the US have to wear adult diapers since they don't get breaks. People may have jobs but, human dignity is at risk. The shocking example from the poultry factory in the US could be the same for AI, especially when it comes to the development and production of AI.Respect for democracy, justice and the rule of law is crucial for human civilisations to flourish. But, AI, bots and algorithms have already been used to rig elections, as was the case in the US in 2016. So, what assurance can the EC give regarding the safety of current and future human civilisations?Principle of Autonomy: Autonomy is already threatened through social media platforms, where algorithms decide what gets promoted and what people see. Followers and Likes are harvested through fake profiles and identities. The right social messages, which are beneficial to people are not disseminated, instead gossip is promoted. Why isn't any relevant global authority able to implement the Principle of Explicability to Facebook? They're acting with impunity. How can citizenry be ensured that this will not be the case in future with more sophisticated AI technologies?The dual-use nature of AI is a recipe for disaster. So much destruction is perpetrated by weapons manufacturers, who are mostly in Europe and the Occident, to has many ramifications.

				sell their weapons for humans to divide themselves, fight and kill each other. This will no doubt continue, albeit in a more aggressive manner, with LAWS. Using AI in weapons systems with a view to reduce collateral is just finding an excuse to produce it despite the unprecedented destruction it can cause. It will also 'play god' and bypass the laws of nature and the philosophical notion of 'Karma', which prevents people from doing bad.			
Katalin	Feher	Budapest Business School University of Applied Sciences	Relevant	Although the cultural aspects has been mentioned (see "cultural sense" on page 7, or in context of demography on pages 9 and 22), the cultural diversity is less emphasised than would be expected. First of all, the AI is a cultural-social phenomenon beyond the technological developments and it is discussed in several academic disciplines with different traditions and values (among others Bloomfiled 2018). Furthermore, it is crucial to implement the cultural diversity as a criteria for AI if the goal is the effective adaptation of technologies in the EU. Last but not least, it would be useful to interpret the AI as a layer on the cultural-social constructions beyond the AI-infrastructure or AI-technology. These perspectives do not just support the frequently cited bias-issue (see mainly in the next chapters), but it is rooted into the techno-cultural perception.	The concept of trustworthy AI seems to be a remarkable approach in context of European values and cultures. Based on the business and IT practice, the question of trusted/trustworthy user or user networks are also suggested to consider in the document. Currently, we have opportunity to understand the process how users learn from machines and vice versa. Before any kind of paradigm shift on this field, it would be a return of investment to understand the motivations and decisions behind the social and cultural habits of the users considering their human network impact. In other words, the term of trusted/trustworthy user is also relevant in the co-operation with machines.	One of the most highlighted concepts of this chapter presents the term of "bias" with relevant approaches. It would be useful to make a difference between the input of bias (which is originated in the human language, culture or society and the machines "translate" them) and the output of bias by an AI system (which is a bias for a system working in context of optimisation, predictive analysis or further methodologies).	Thank you for the opportunity of reading and commenting the meaningful document what has been attached. It would be grateful for me to work also on it in next phases.
Krzysztof	Potempa	BRAINCURES LTD	I do not agree with the statement in section (iv): "AI can help humans to identify their biases, and assist them in making less biased decisions". AI is only as good as the best medical experts algorithm to diagnosing retinal disease (see PMID: 30104768). Consequently, AI is only as good as the human-rules that have been programmed into it...in the medical space I have yet to see a machine become better than the master who programmed it. This masters knowledge embedded into a support system	1. It would be great to expand more on section A point (i) concerning increasing public and private investment in AI to boost its uptake 2. The black box nature of most AI systems may make it difficult to implement section 3.3 provision for a due process by design, meaning a right to a human-centric appeal, review and/or scrutiny of decisions made by AI systems 3. It may be very difficult to make sure that AI systems are developed and implemented	1. Really like the statement that AI applications should not have a one-size-fits all approach, but be user-centric and consider the whole range of human abilities, skills and requirements (p15 maybe make it bold) 2. Consider to mark in bold: Machine learning algorithms identify patterns or regularities in data, and will therefore follow the patterns resulting from biased and/or incomplete data (p16, Non-Discrimination section). This suggest the need for the	1. Consider in bold: Have the limitations of the product been specific to its users (p27) 2. Healthcare Diagnose and Treatment assessment list for companies that have a steady framework of genes that power decision making and not just black box algorithms, it should be possible to benchmark potential retrospective power of the algorithm. Investment will then be needed to take perform prospective testing. As an example, the BRAINCURES Discovery Engine is a framework comprised of 822	

			may though help more junior persons perform better.	in a way that protects societies from ideological polarization and algorithmic determinism (just think of what items are selecting if Alexa is helping you to shop) p9	collection of robust data sets that then allow us to harness the full potential of AI in specific domains. 3. Consider to mark in bold: the complexity, non-determinism and opacity of many AI systems, together with sensitivity to training/model building conditions, can make it difficult to reproduce results (p17-8-Robustness) 4. It is important to realize that AI the medical domain is only at the learning stage from haystacks of limited quality data and produces outputs that are usually not very understandable 5. Would be great to make a figure that captures the essence of Trustworthy AI, i.e., that a AI system's capabilities and limitations are provided in a clear and proactive manner, being realisitic, traceable, auditable (p23). Take BenevolentAI and IBM Watson as two examples in healthcare where the communicated ability of the systems is quite a futuristic one.	target genes that can be used to de-risk and accelerate drug discovery. For example, our knowledge raises the Phase III success rate of kinases from 62.5% (i.e., 6 kinases failed, 10 kinases succeeded) to 75% (i.e., 1 kinase failed, 3 kinases succeeded) by excluding 12 of the 16 kinases from a benchmark set of 331 genes with Phase III outcomes. For more details please refer to: https://www.linkedin.com/company/braincures/	
Monique	Calisti	Martel Innovate	An opening, clear and crispy statement about what is the purpose of this document is somehow missing. Do you want to raise awareness? Do you want to initiate some specific policy debate? Do you want (as I think you do) engage a growing number of researchers and innovators to ensure enforcement of ethical AI across EU and beyond? Then say that more explicitly from the very beginning.Is there any description of the next iterations? Is the "final" version in March planned to be made available to the broad audience or discussed with MEPs or inject in any other more specific discussion about MFF? I think this should be explained in this section.From a technical point of view (not political) does it make sense to talk about a European approach to AI? Is it enough to ensure that in Europe we "produce" and "stick" to "trustworthy AI" to reap AI's benefits in Europe? I would suggest trying to make more explicit what is the expected impact/value of building a high-level framework.I find it misleading, almost disturbing that you say these guidelines "..aim to use ethics as inspiration to develop a unique brand of AI". We should aim at ethics becoming a core pillar of of any AI-based system and that AI per se is developed in an ethical way.	Figure 2 could be replaced by a specific one. More interesting would be a picture showing what are the core pillars for the trustworthy AI framework these guidelines are setting - how the rights (section 3) are interleaved with the principles and values (section 4).Human centricity for the AI made in Europe would then mean to follow the ethical principles and values as defined in this chapter right?I am not sure about whether to have an "ethical" expert for to accompany the "design, development and deployment of AI" is anything realistic. AI is already here, it's already deployed and I doubt, especially at development and deployment level, to have a ethical expert would address the need to verify whether an AI system behaves indeed ethically. Not sure how an ethics expert could deal with code and algorithms. Let's face it. If we don't have ways to verify and control from a technical point of view what's going on with AI in action, it's very difficult to judge on whether its behavior is sticking to given ethical principles. So it boils dow to educate people and provide them with strong ethical values that indeed can be the ones you identified. So my major concern is that while some will develop AI, many more will use AI off-the-shelves without even understanding fully its intrinsic functioning...unless they do have AI experts on board. So I would be very clear at discussing and setting a very high priority for all of us in the future: how can we ensure to teach and develop AI skills so that we will be in control of AI deployments?About critical concerns see my point above. I'd say a very critical aspect about use and deployment of AI is in creating the proper skills to understand and master the AI reasoning/learning mechanisms behind it.Finally, to me ethical AI must account also for sustainability, like any other technology we develop and deploy.NOn technical methods: education and development of AI skills is to me a core element. Most digital businesses will have to face within the next couple of years how to grow AI skills/expertise in-house.	I am not sure t trustworthy AI goes together with "design for all"? I am not convinced. Depending on what is the purpose of any deployment of AI the design must match the requirements and needs of the specific targeted users. So clearly there will NOT be a "one-size-fits-all" approach that to me means indeed you cannot design an AI system for all. Depending on the application, it must be obviously possible to have to include "all" in the society including minorities or people with disabilities. So maybe better to talk about "INCLUSION" rather than "DESIGN FOR ALL" to express this concept.About "respect for human autonomy" I'd be more concerned about educating people to critical thinking and critical behaviour. Technology is obviously not neutral, so we must ensure regulation and education can help citizens in having a more conscious and aware relationship with technology.	Accountability is not an easy piece of cake as this will finally indicate who pays the bill when something goes wrong - see driverless cars. That's maybe why so many insurances are investing so much on analysing AI deployments.Especially at this level, I believe to develop AI expertise - to be able to verify its functioning - is crucial to any public and/or private institution. Understanding the use and implications of AI usage though might not be something which can always be done in a short period of time. Which is why validation and testing in the R&D life-cycle should / will play a more crucial part, especially when deploying AI for critical tasks.	Try to be very explicit about WHY do we need to talk about ethical AI: what is different than ethical programming / ethics of programming?For the document to be more impactful: more infographics and concrete data / examples. This would help teh readers to better identify themselves into this initiative.

Florian	DAMAS	Nokia	Nokia welcomes the draft guidelines and wishes to thank the distinguished members of the HLEG for their efforts and guidance, given the complexity of the topic. We also thank the HLEG and the European Commission for the opportunity to comment on these draft guidelines, as a multi-stakeholder dialogue and approach to the development of this type of document is crucial for its legitimacy, wide-spread adoption and implementation. We recommend that the same approach be used for the upcoming documents produced by the HLEG, including the guidance related to use of AI by public authorities, to ensure democratic participation, especially in light of the widespread potential uses of AI in government and society. We believe such guidelines are needed for AI to be accepted as a trustworthy technology (i.e. where it is generally understood that its benefits outweigh its risks'). Also, such guidelines provide a governance framework for the development and use of AI technologies across Europe, for reducing the barriers of AI development in ethical ways, and to reach alignment and gain support from the industry. We recommend a regular evaluation and review of the guidelines, to allow for the adaptation of these based on knowledge and technological advances in AI, as well as the anticipated evolution of concepts such as 'privacy', 'human values' and 'fundamental rights. We applaud the efforts to formulate 'critical concerns', while phasing out unfounded worries and misconceptions and to launch an informed public discussion on these, considering the high stakes. We would like to stress the importance of a dynamic and competitive market of AI. We are seeing an increase in concentration of market power among a relative few companies in the data economy sector and in industries where digital transformation is under way, something that more and more SMEs, governments and citizens are coming to rely upon. These companies are already expanding their market power thanks to AI. Policymakers might seek to create the technical and legal conditions to facilitate data sharing between companies and public administrations to train Artificial Intelligence applications, and sector-specific measures in order to enable fair and open competition. We recommend that the guidelines place increased emphasis on the need for a risk-based approach during the early development phases of AI, in which the developer evaluates the potential impact on human beings by reference to the intended use of the AI system. We recommend that the purpose of these guidelines is clarified from the perspective of their acceptance or endorsement by those who create, commercialise or use AI systems. While we understand the current desire to maintain the voluntary character of these guidelines, there are numerous references in the document to "compliance" and "requirements". Should the final version of the guidelines include a specific voluntary endorsement mechanism, we believe that	Ethics by design should be applied in a differentiated manner, depending on the use cases and the user groups on which AI will have an impact (a 'user group' being defined as a group of users sharing the same or similar cultural habits, values and customs). European values and human rights provide a good reference basis for assessing ethical design, but a multi-stakeholder approach should be employed, given the cultural diversity of Europe, as well as the borderless character of AI and technology in general. Between UN and European human rights standards and instruments, it is unclear which should be the acceptable reference. Gaps still exist currently between the range of human rights recognised under EU human rights law, by countries member of the Council of Europe and by the UN human rights instruments to which the Member States are party. There are also differences between the depth of the obligation to guarantee human rights under UN instruments, under the European Convention on Human Rights, under the European Social Charter and under national legislation. We recommend the HLEG to investigate further on this topic. Nokia would prefer the standards and instruments to be set at a global rather than regional level, with strong accountability mechanisms that are established and widely recognized. For this purpose, a few questions could be considered: (1) Can the EU afford to limit its approach to the European perspective, especially if its ambition is to have a positive impact on ethical standards globally? (2) Is it acceptable that other countries using or creating AI systems are not applying ethnical principles? (3) Is it acceptable that certain countries employ different AI governance systems than the European Union Member States, which might be based on interests that may be considered incompatible with fundamental human rights and values?	We would like Governments to launch a call for action specifically targeted on AI as a tool for innovation: 1. to provide the necessary skills and education required for citizens and businesses to understand AI; 2. to prevent and mitigate potential malicious use of AI (examples: manipulation, fake news…); 3. to convey recommendations on key enablers for different sectors to benefit from AI. The policymaker also has a role also to ensure a broader enlightenment of the general public of what AI is. That the often-unspecified anxieties and unease are generally unfounded, because AI is in many aspects rather an evolution which has already started many years ago with innovative coding and algorithms. Also, AI is not generally used to replace but rather to enhance human capabilities and decision-making, and AI-related risks are controllable. An ethical framework should also serve this purpose, i.e. giving people confidence in the use of such technology, based on sound ethical standards, so as not to feel intimidated by it. In general, policymakers should try to achieve the right balance between the 'precautionary principle' and the 'innovation principle'. This aspiration applies equally to the policy area of AI. Before diving into the Principles, we note that some existing regulations should be adapted to ensure that barriers to the development of AI are eliminated. In certain contexts, current or prospective legal instruments (such as the GDPR, the draft ePrivacy regulation) are not as conducive to innovation as they could be and there is great potential for them to be misinterpreted or applied in a manner that inhibits the sharing of information and the free flow of data, under the disguise of privacy or national security. On the principle of Explicability of AI decisions (page 10) It is currently unclear how this principle should be interpreted, and which levels of transparency and accountability should be required. Which interpretation will lead to desired policy outcomes? Also, the results when applying AI may not always be explicable (this is inherent to the complex nature of most AI applications). Humans may not be capable of understanding those results. Does this exclude the application of the Principle of Explicability, as its application could impede the benefits of AI? Introducing the Principle of Explicability presents the risk of not being able to use AI where it is most suitable: to reach decisions in the most complex cases of a multitude of interrelated categories of input data. Should the focus rather lie on the implementation of other safeguards, to ensure correct and verifiable outcomes, especially in the case recursive AI, for which human control would be foreseen)? This could also have an impact on the discussions about assignment	Ethical impact assessments (used alongside privacy impact assessments) are valuable instruments and we welcome the introduction of the concept of an "Assessment List" in Section III. The questions are pertinent. However, for this list to be actionable and truly valuable (especially to those who lack the resources to conduct assessments based on expensive human rights due diligence corporate programs), we suggest that it be expanded with recommendations or solutions to address certain issues uncovered by going through the assessment list. General pointers such as the "Establishment of an Ethical Review Board" may be appropriate, but in reality, these may not necessarily be helpful to SMEs who lack the required resources. Another important remark is related to the fact that this assessment is focused mainly on AI that uses personal data or that is used in relation to consumers. The assessment should also provide solutions that are proportional to the problems that the AI is attempting to solve versus social impacts, versus risks of malfunctioning. Ideally, the assessment should provide companies with the necessary means to assess the fitness of an AI system for deployment, which would make this an extremely useful instrument and a true enabler for AI, while also setting the basis for formal assessments, endorsements and even certifications.	Glossary We also recommend the use as much as possible of generally accepted terminology and of legal concepts used elsewhere in European legislation, rather than the introduction of new terms. Also, we recommend expanding the Glossary substantially with terms currently explained in footnotes. In relation to this, it could be useful to also add a list of reference materials (currently also available via the footnotes). We acknowledge the difficulty in defining Artificial Intelligence. The definition provided is very detailed, but it seems to point towards a narrower field of application than other existing definitions, which could result in the exclusion of certain systems from the scope of applicability of the guidelines. The definition seems to focus on systems enabled to take decisions, while excluding systems which generate certain findings or merely suggest the best options for resolving problems, while decision-making remains the prerogative of human beings. If decision-making is seen as an important element of the definition of Artificial Intelligence, we suggest providing further guidance on how to verify whether decisions taken by AI are ethical/just/unbiased or not.

the legal consequences of this endorsement should be carefully considered, especially in light of existing European legislation in relation to article 6 of the EU directive 2005/29/EC on unfair business-to-consumer commercial practices.

of liability and legal personality.

On the critical concerns (from page 11)

The focus of the draft Guidelines seems to be on AI that uses personal data, hence the many references to informed consent of data subjects. It is important to also note that personal data may not be the only, or main category of data used by AI, particularly in a B2B context.

In relation to personal data in particular, we feel compelled to point out that while user choice should play a central role in some AI deployments, a significant number of AI use cases may exist where personal data is processed validly and lawfully, while based on legal grounds different than consent, as foreseen in the General Data Protection Regulation: 'legitimate interest', 'further compatible processing', 'vital interest', etc.

Further, we recommend that the ethical and human rights assessment of AI systems should take into account and build upon the privacy impact assessment required under GDPR rather than foreseeing an additional privacy assessment mechanism. From that perspective, we also recommend that the guidelines should focus more on methods to be transparent about the way personal data is used in an AI context, increasing awareness about the impact of AI on privacy while also providing the reassurance that AI systems that follow the guidelines are trustworthy and beneficial to individuals and to society.

On covert AI systems, it is important for citizens and businesses to be able to rely on human-based decision-making if AI service objectives are not met.

On scoring, we question the need for proper rules on data handling as it can be as harmful without AI.

On Lethal Autonomous Weapon Systems (LAWS), there are severe concerns regarding the usage of AI for military purposes. We believe that simply applying the rule of law may not be the best option. Something legal may not be ethical at all times and vice versa: what may be considered acceptable on the human right side may not be adequate for security reasons. Ethics by design should then take the upper hand.

On the technical methods (from page 19)

We support the decision to list technical and non-technical methods for achieving trustworthy AI and we recognize that there are actions being undertaken to develop these methods.

Claudio	Telmon	Clusit / Copernicani	Goals of AIThe adopted definition of AI is "Artificial intelligence (AI) refers to systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take (according to pre-defined parameters) to achieve the given goal. AI systems can also be designed to learn to adapt their behaviour by analysing how the environment is affected by their previous actions". This definition may be correct for AI, but not for systems implementing AI. Systems implementing AI may have different goals, depending e.g. on other market requirements, and these different requirements may explicitly limit the ability of the system to "achieve the given goal". Product goals are shaped by the market and are often quite different from the models they started from. An example seems to be the adoption of watermarking techniques in AI. These techniques embed some specific data in the AI system (e.g. by training it on the data), that can be detected afterwards (see e.g https://www.ibm.com/blogs/research/2018/07/ai-watermarking/) and that cannot be easily removed. The goal is protecting intellectual property, and not achieving the AI goal. These techniques may slightly alter the results and therefore the effectiveness of the AI in achieving its goal. While these differences may be limited, two issues arise: evaluations on the impact of these techniques seem to be empiric somehow, maybe we should therefore ask if the safeguards are enough: should any of these techniques have a relevant impact on the effectiveness of the AI decision process, should we accept it? when an error occurs, it may be unclear if it is due to the proper AI behaviour or to the bias voluntarily added by inserting the watermarking; what is the impact of this uncertainty e.g. on redress issues?Of course, this is just an example: market and service providers interests and needs may significantly move the goal of AI from its basic definition, causing less-than-optimal decisions to be common. This issue should be explicitly dealt with.			Section III. Assessing Trustworthy AIThis is a minor issue. Assessments usually provide some kind of "measurable" result on the assessed issue. So, questions usually are in a form that can provide a yes/no, high/low, compliant/non compliant answer. The current list, while very fit for identifying relevant issues and often written in the proper way, is sometimes not providing a clear answer.As an example:"Who is accountable if things go wrong?"Questions may be something like:"Has the accountable subject been identified?""Has the accountable subject the power and resources to deal with e.g. redress issues?"and so on.Also, being accountability a critical issue, one question should be something like:"Is the accountable subject able to demonstrate compliance with applicable laws?" etc., see e.g. GDPR art. 5.2	1. Risk based approach. White risks are frequently cited in the document (e.g by stating that AI adoption should "maximise the benefits of AI while minimising its risks"), a risk-based approach is not defined nor explicitly considered, while an assessment checklist is proposed at the end of the document. A risk based approach may fit the "technical methods" described e.g. in section 2, but it may be useful to be much more explicit on this. While the document overall set several relevant safeguards as requirements for Trustworthy AI, these requirements may be hard to properly implement and evaluate if risks are not explicitly identified as part of the process.A risk-based approach is adopted e.g. by the GDPR as a founding principle of the overall approach, not just as a technical implementation, and it seems to be quite effective from these first months of adoption, especially within the Data Protection Impact Assessment (DPIA) as defined in art. 35. This risk-based approach is consistent with best practices when dealing with possible impacts that cannot be already fully foreseen and evaluated. Also, being already widely adopted by GDPR, companies and citizens are being used to the required mindset, and procedures are being defined, so that a limited effort may be required to extend this approach to other cases, such as AI adoption. In fact, GDPR provisions may already cover several critical cases for AI adoption. This highlights another relevant issue: what are the risks actually discussed in the paper when stating that they should be minimized? While there are examples on use cases and general assertions, no clear identification of these risks seems to be described, so a general risk identification and management approach should be adopted. Again, GDPR clearly deals with risks "for the rights and freedoms of natural persons", which is a very wide definition, but are these the only risks we are dealing with AI? Should other risks be considered, e.g. risks for the community as a whole, and not just for individuals? Without any clear identification of these risks, most of the discussion is hard to be effective besides on the discussed use cases and on similar cases.The document should also consider recommanding to have a continuous monitoring of emerging topics that may need a specific regulation, e.g. because of high risks involved in the research or adopion of AI solutions in that areas.2. Accountability and sanctionsGDPR makes the data processor accountable for its choices and actions, limiting authorization procedures to very specific cases. The effectiveness of the GDPR depends therefore on the relevant sanctions that may be imposed to data processors. A similar attitude should at least be considered. Otherwise, market rules that are already damaging the IoT market with insecure components, could also damage the ai market, by providing cheap but insecure or otherwise dangerous AI services and components. We are not convinced that voluntary adoption of codes of conduct would have any impact on this kind of issues. Consider some cheap "AI processor" in a smartphone, would anything but imposing accountability and high fines have any impact on the development, marketing and adoption of this component? We fear not. It must also be noted that enforcement,

								more than regulation itself, is often the weak point in ensuring accountability. For example, even the requirements already set forth in Art. 22 of the GDPR do not seem to be fully implemented by many platforms that ordinarily use AI algorithms to evaluate human action (for example, automated content review on social networks, and possibly automated upload filters if adopted in the future). Some specific effort should be taken to ensure that any legal requirements are actually enforced.3. Non-EU AI products and servicesThe document deals with "Trustworthy AI made in Europe". This is quite different from ""Trustworthy AI adopted in Europe". Should we expect that any ethical constraint defined for "made in Europe" will have any impact, besides limiting the market share of EU-made AI? History of many technologies shows that cost-effectiveness and several other drivers are far more relevant than security or safety issues. Again, consider the IoT, or smartphones markets: components are mass-produced in the Far East, and no European regulation seems to have any impact on their features and quality. In fact, we can already assume that a relevant part (most?) of AI solutions adopted in Europe won't be "made in Europe". In fact, a much stronger position is currently being proposed at the EU level for cybersecurity risks stemming from low quality IoT, see https://ec.europa.eu/info/law/better-regulation/initiatives/ares-2017-3436811_enGDPR on the other hand, seems to be effective also when dealing with foreign countries offering services to EU citizens. This is a critical issue, that may be more explicitly dealt with in the document. Setting any limit only to the European AI initiatives may hinder the ability of EU companies to compete, without actually protecting the European market and citizens. On the other hand, setting the same limits to companies offering services to EU companies and citizens would spread the European ethical approach to other countries. While this may be a policy issue to be dealt with in future documents, defining the context of the EU action seems to start from this document.
Anett	Molnar	HIMSS / PCHA	The comments below are submitted jointly by Healthcare Information and Management Systems Society (HIMSS) and the Personal Connected Health Alliance (PCHA), a HIMSS Innovation Company. As a key not-for-profit group in the healthcare IT and digital health sector, we represent the interests of healthcare provider organisations, healthcare payers, healthcare professionals and patients. Through our members and the services we provide we have a clear insight into the opportunities and challenges of using artificial intelligence across the healthcare chain, from drug discover to care provision, at all times underpinned by patient empowerment and equitable access to care. HIMSS and PCHA warmly welcome the initiative of the High Level Expert Group and recognise the value of developing clear, understandable and user-friendly ethical guidelines for the use of AI as an emerging and evolving tool with power to change the lives of European citizens across a wide range of sectors and industries. The	HIMSS welcomes the initiative of the European Commission in bringing together a High Level Expert Group on AI. We read the Draft Guidelines with great interest and value the focus on the overarching ethical questions posed by this powerful new technology. HIMSS welcomes the inclusion of healthcare in the list of areas noted in the Draft Guidelines; we would argue however that a wider focus that diagnosis and treatment would be appropriate, and would urge that in future iterations the Guidelines should include concepts of health, which embrace well-being and disease prevention. HIMSS suggests that a wider construction of health would better reflect the promotion of access to health, well-being and healthcare as reflected in historical human rights statements, which are cited as the basis for the rational of the Guidelines. Of particular note here is article 12 of International Covenant on Economic, Social and Cultural Rights (1966) which states (inter alia) that	HIMSS welcomes the ten key requirements set out in Chapter II, all of which have a high level of importance in the use of AI in healthcare. One particular element that could be expanded is the concept of Human Oversight to include the rights of an individual to exclude unwanted actions. It should be noted that this concept is reflected in the case law of the European Convention on Human Rights, which forms part of the ethical base-line of European policy. The case law has established a wide interpretation of the right to respect for private life guaranteed by Article 8 of the Convention as covering the right to refuse medical treatment or to request a particular form of medical treatment (Glass v. the United Kingdom; Tysiąc v. Poland). It is important to note therefore that the use of AI in healthcare, or indeed any other sector, should retain as far as possible the right for an individual to refuse a particular treatment or action. A key step to achieving this could be by ensuring that the patient voice has a	HIMSS welcomes the intention to use Healthcare Diagnose and Treatment as one of the four use-cases and tailored assessment lists. As noted above, we would argue that the description should be expanded to include wider concepts of health, notably disease prevention, well-being and mental health. Although mentioned in passing, it is worth noting that maximising the impact of AI in healthcare will require adequate and targeted education on AI for healthcare professionals, in particular to equip the healthcare workforce with a good understanding of AI empowered technology and proper training for interpretation of the data presented by the systems assisted by AI. HIMSS endorses the issues and areas included in the outline assessment list. In developing such an assessment list specifically for the use of AI in health, it will be important to include health specific	As for future steps, HIMSS recognises the importance of the AI ethics guidelines to be acknowledged in best practices of procurement of AI enabled IT systems. HIMSS would be delighted to draw on its wealth of experience in the use of new technologies in healthcare to supporting the development of a healthcare domain specific assessment list for ethical use of AI in healthcare.

			comments submitted below focused on the role of AI in health promotion, care provision and the objective of driving sustainable and inclusive health systems for all.	States shall recognize the right of everyone to the enjoyment of the highest attainable standard of physical and mental health. HIMSS fully endorses the focus on core ethical concepts and notes that many of these intersect perfectly with core concepts of beneficence and non-maleficence, equality and access enshrined in medical ethics and medical law across the EU. We would urge the Group however to grapple further with some of the potential challenges which the use of AI in healthcare can raise, notably the potential to predict future health states of individuals, including for disease areas in which there are currently few available treatments, such as dementia and Alzheimer's disease. A particular area of focus of AI in healthcare should include public health, which comprises both population health measures as well as personalised interventions. The power of AI to drive new approaches in public health has been noted, but this in turn raises questions about the ethics of secondary use of data, which require further exploration both in legal guidelines and practical solutions. At HIMSS, we would argue that new models of dynamic consent should be a core tool to develop more patient centric and research friendly models of informed consent, as foreseen in framework legislation such as the GDPR.	place in the stakeholder dialogue, and the inclusion of patients in diverse design teams. On a practical level of realising trustworthy AI, the role of interoperability between systems built on AI must not be under-estimated. HIMSS would urge the HLEG to include reference to interoperability as a tool for ensuring that core ethical values such as privacy, safety and transparency can be achieved. The concept of interoperability is particularly important in the field of healthcare, where data obtained from a wide range of sources and incorporate many different types of data sets. Interoperability between those data sets is a key tool in driving trust between the players in healthcare and overcoming the shortcomings of siloed data. HIMSS sees the huge potential of AI in healthcare, but would argue that without due attention to interoperability of AI solutions and approaches, that potential will be hard to achieve.	needs, including but not limited to: • Accountability - training and support for health care professionals and patients in understanding the power and capacity of AI in healthcare and how it is /may be used • Data Governance - data stewardship as a core component of data governance for Design for all – inclusion of end users, both care providers and patients, in design • Non-discrimination - the need to guard against the identification of risk groups who could be disenfranchised from medical care • Privacy - the need to balance the ethical principle of privacy and the interests of an individual in privacy, with the public health benefits of health data as a public good which can be used to drive research for better care, new medicines and new treatment models. As AI applications in healthcare make use of many different kinds of data, many of which will be legally considered as sensitive and therefore will benefit from a certain amount of privileged legal protection. However, much of the potential of AI in healthcare arises from bringing non-sensitive data into partnership with medical (sensitive) data, such as social media activity and internet search history. Special and new ethical approaches are needed to ensure that when these two categories of data are used jointly, this is done with due respect for the information about the health of the data subject which could be revealed without the data subject having full transparency on this capacity.
Anonymous	Anonymous	Anonymous	Please clarify who is the guideline for? Developers of AI? and it is aimed at protection of people who are using AI or AI itself? do we consider AI who might also have feelings and compassion subjects for protection?				
Anonymous	Anonymous	Anonymous	Under the following section, some supplementary relevant information should be provided."trust in the rules, laws and norms that govern AI – it should be noted that no legal vacuum currently exists, as Europe already has regulation in place that applies to AI – or trust in the business and public governance models of AI services, products and manufacturers"It says Europe already has regulation in place that applies to AI. Please provide an appendix of what laws exist that apply to AI. e.g. GDPR and what else and demonstrate how it relates in real world practice?e.g. GDPR - covers data privacy, but how about the misuse of data we have given permission to be shared but not in the way were informed?The appendix should say...GDPR laws cover data privacy etc.Xyz laws cover GAN deep fake images...Xyz laws cover AI assisted hackingXyz laws cover big data management and sharing etc.Please give example legal cases / precendents.	Recent advances in GAN Tech mean it is possible to easily create DeepFakes. Although the technology alone is not harmful there is a high risk it would be used to slander or bring individuals to disrepute. The viral nature of social media means the damage can be irreversible.I propose a guideline that mandates that all Deep Fake output must be identified as such on a permanent watermark or label in a video or image. It should be overlayed and non-removable. This is to prevent social upheaval due to misinformation from fake news outlets or individuals with an agenda to mislead.Alternatively at the very minimum AI providers should watermark every deep fake face swap so social media outlets can easily identify and label these videos and images as such.	Please include a paragraph that...- Describes how slowing down the pace of deployment can mean a better product... E.g. Move fast and break things Vs an ethical rollout of AI limiting bias - Describe the benefits of large big data firms should make open anonymised datasets to allow AI to advance and tech firms to compete- Why we should avoid a winner take all scenario - that takes up all the AI talent and no diversity in AI and ultimately harms innovation.	Machine Learning algorithms used in a public scenario should be evaluated and certified every year that it is fit for purpose because Machine Learning models can decay and lose accuracy over time.In Electrical Engineering machines must be calibrated annually and certified they are calibrated in high precision use cases e.g. healthcare, manufacturing and mission-critical machines. The calibration certificate is issued by an independently certified tester and the private company pays to have this service performed. If it fails the test, the vendor company will pay the independent company to recalibrate to bring it to the legal standard.This certification is financed by vendor companies but ensures hardware is up to the standard or the product cannot be sold or legally be used. Now transfer that idea to Machine Learning algorithms. If upon evaluation it is determined there is a deficiency, the Machine Learning Calibrator cannot issue a certificate. In other words, there needs to be a Machine Learning Calibration performed regularly to ensure algorithms have not lost effectiveness, have not been manipulated and have not absorbed bias from a given dataset. It is done by an Independent company so that the vendor cannot fudge, or avoid addressing the issue.This is especially important in government Machine Learning	I am supportive of the work and commend the guidelines. Please consider my suggestions for additions below...Recent advances in GAN Tech mean it is possible to easily create Deep Fakes. Although the technology alone is not harmful there is a high risk it would be used to slander or bring individuals to disrepute.The viral nature of social media means the damage can be irreversible.Deep Fakes are now becoming so convincing that it's hard to tell when videos have been altered.This has serious ramifications for Political situations- Where people with an agenda intended to manipulate public opinion and affect votes- It could cause social upheaval where a person is quoted as something controversial they did not sayMy suggested questions or topics of concern are on education, regulation and prevention of misuse... - Shouldn't any AI vendor that supplies Deep Fake tech, be legally required to supply an algorithm that people can use to detect it as fake also? - What can be done to Identify explicitly when a video or photo or voice has been faked? - Should all Videos / Image have a non-removable label to indicate fakery? - Should all mimicked voices always have an audible fingerprint that can be easily identified as AI generated? - Should AI providers watermark every deep fake face swap so social media outlets can

						algorithms used to decide who will be issued housing benefits, welfare, voice recognition and facial detection. Society needs to know checks and balances are in-place where the public is impacted. There must be an independent certification and remediation system we can put our trust in.	easily identify and label these videos and images as such? - Should Companies be mandated to explicitly watermark every deep fake video to indicate it is a fake and not be taken seriously? - Should there be a digital fingerprint on every deep fake video produced to trace who produced such a video (for accountability and deter misuse)Please do pass on these questions for consideration by the Panel.
Jon Toivo	Hansen	Independent legal researcher, LL.M.	No suggestions.	In accordance with the proposition that AI ethics should be based on the fundamental rights of citizens, it is good form to clarify the reach of the duty to protect such rights. I recommend the inclusion of text (here and in the following) to that effect, e.g.: Ad I.1. "The EU's Rights' Based Approach to AI Ethics" (p. 5) After the fourth paragraph ("[…] fundamental rights endorsed by humans."), I suggest including: "As the rights underlying the ethical principles and values are indeed fundamental, they are to be universally observed and protected. Consequently, anyone benefiting from the development, deployment and use of an AI system shares in the responsibility for the ethical operation of that system."	Ad II.1.1 "Accountability" (p. 14) I suggest rephrasing the section: "1. Accountability Good AI governance presupposes the highest possible level of accountability. Those that benefit from the development, deployment or use of AI systems should consider and clarify the relationships of accountability and legal liability towards other stakeholders, such as partners (e.g. customers, suppliers, producers), regulatory authorities, and industry bodies. Lawmakers should make sure that the rules governing AI are shaped to avoid ambiguity and blame-shifting, especially with regards to consumers, employees, and citizens acting in their non-professional capacity. Anyone should be entitled to sufficient information and effective redress from an entity that has facilitated their contact with an AI system." I furthermore suggest that the remaining discussion of mechanisms be relocated to and incorporated into II.2.2(1st bullet) "Regulation" (p. 21), thus replacing that section with the following text: "In their regulation of AI lawmakers should consider the diversity of available accountability mechanisms, and apply the method of redress best suited to ameliorate any negative consequence of the operation of AI systems. Mechanisms can range from criminal jurisdiction over strict liability or culpability following compensation for damages to rectification or reconciliation without monetary relief. The designated accountability mechanisms should also take into account the nature and weight of the activity, as well as the level of autonomy at play. For example, an instance in which a system misreads a claim for medical expenses and wrongly decides not to reimburse may be compensated for with money. In a case of discrimination, however, an explanation and apology might be at least as important. Regulation should also at least require anyone benefiting from the development, deployment, and use of AI to expressly	Ad III.1. "Accountability" (p. 24) Add to 1st bullet: "[Who is accountable if things go wrong?] Have all AI systems and responsibility relationships been clarified, considered, and properly mapped?" Ad III.1. "Transparency" (p. 27) Reword 4th and add 5th bullet: "Have the remaining criteria for deployment of the AI system been set?" "Has the user been informed of the AI system's purpose, criteria, and relationships, as well as his/her rights?"	

					document and disclose to which systems and for what purposes they are subjecting a person or their data and inform them of their rights. Specific considerations of revision, adaptation, or introduction of regulation, such as safety legislation or liability frameworks, are beyond the scope of the Ethics Guidelines for Trustworthy AI. Such policy recommendations are the subject of discussion in the AI HLEG's next deliverable."			
Anonymous	Anonymous	Anonymous	Under the following section, some supplementary relevant information should be provided."trust in the rules, laws and norms that govern AI – it should be noted that no legal vacuum currently exists, as Europe already has regulation in place that applies to AI – or trust in the business and public governance models of AI services, products and manufacturers"It says Europe already has regulation in place that applies to AI. Please provide an appendix of what laws exist that apply to AI. e.g. GDPR and what else and demonstrate how it relates in real world practice?e.g. GDPR - covers data privacy, but how about the misuse of data we have given permission to be shared but not in the way were informed?The appendix should say...GDPR laws cover data privacy etc.Xyz laws cover GAN deep fake images...Xyz laws cover AI assisted hackingXyz laws cover big data management and sharing etc.Please give example legal cases / precedents.	- When a company shares your data as in a way that was not agreed originally, then should the company reach out again for renewed permissions?- When collecting data companies should provide informed consent of exactly how the data will be used e.g. bullet-pointed in simple lay-terms and not hidden behind legal lexicon.- Video sharing sites like YouTube should automatically detect if a video contains a GAN generated video. It should automatically LABEL that this video is GAN generated so viewers are not easy mislead. - ADULT X-Rated websites should NOT allow deep fake image to be uploaded as this is harmful to the women who have had their faces transferred without their permission. This is possible to do with the technology available today, there needs to be the willpower to enforce it.Recent advances in GAN Tech mean it is possible to easily create DeepFakes. Although the technology alone is not harmful there is a high risk it would be used to slander or bring individuals to disrepute. The viral nature of social media means the damage can be irreversible.I propose a guideline that mandates that all Deep Fake output must be identified as such on a permanent watermark or label in a video or image. It should be overlayed and non-removable. This is to prevent social upheaval due to misinformation from fake news outlets or individuals with an agenda to mislead.Alternatively at the very minimum AI providers should watermark every deep fake face swap so social media outlets can easily identify and label these videos and images as such.	Please include a paragraph that...- Describes how slowing down the pace of deployment can mean a better product... E.g. Move fast and break things Vs an ethical rollout of AI limiting bias - Describe the benefits of large big data firms should make open anonymised datasets to allow AI to advance and tech firms to compete- Why we should avoid a winner take all scenario - that takes up all the AI talent and no diversity in AI and ultimately harms innovation.	- When working out AI Bias, what techniques are most popular for reducing them? - How do people measure the impact of tradeoffs? - how many companies are transparent about what tradeoffs have been made? - Should every model come with a statement of its limitations for end users?- Does the algorithm have a feedback loop to correct and mitigate errors?From the perspective of the end user, they may become complacent and unquestioning of a bad decision. If the company is transparent of where it is deficient (e.g. scoring ethnic faces) the user can be vigilant and use their own superior judgement on occasion.Machine Learning algorithms used in a public scenario should be evaluated and certified every year that it is fit for purpose because Machine Learning models can decay and lose accuracy over time.In Electrical Engineering machines must be calibrated annually and certified they are calibrated in high precision use cases e.g. healthcare, manufacturing and mission-critical machines. The calibration certificate is issued by an independently certified tester and the private company pays to have this service performed. If it fails the test, the vendor company will pay the independent company to recalibrate to bring it to the legal standard.This certification is financed by vendor companies but ensures hardware is up to the standard or the product cannot be sold or legally be used. Now transfer that idea to Machine Learning algorithms. If upon evaluation it is determined there is a deficiency, the Machine Learning Calibrator cannot issue a certificate. In other words, there needs to be a Machine Learning Calibration performed regularly to ensure algorithms have not lost effectiveness, have not been manipulated and have not absorbed bias from a given dataset. It is done by an Independent company so that the vendor cannot fudge, or avoid addressing the issue.This is especially important in government Machine Learning algorithms used to decide who will be issued housing benefits, welfare, voice recognition and facial detection. Society needs to know checks and balances are in-place where the public is impacted. There must be an independent certification and remediation system we can put our trust in.	I am supportive of the work and commend the guidelines. Please consider my suggestions for additions below...Recent advances in GAN Tech mean it is possible to easily create Deep Fakes. Although the technology alone is not harmful there is a high risk it would be used to slander or bring individuals to disrepute.The viral nature of social media means the damage can be irreversible.Deep Fakes are now becoming so convincing that it's hard to tell when videos have been altered.This has serious ramifications for Political situations- Where people with an agenda intended to manipulate public opinion and affect votes- It could cause social upheaval where a person is quoted as something controversial they did not sayMy suggested questions or topics of concern are on education, regulation and prevention of misuse... - Shouldn't any AI vendor that supplies Deep Fake tech, be legally required to supply an algorithm that people can use to detect it as fake also? - What can be done to Identify explicitly when a video or photo or voice has been faked? - Should all Videos / Image have a non-removable label to indicate fakery? - Should all mimicked voices always have an audible fingerprint that can be easily identified as AI generated? - Should AI providers watermark every deep fake face swap so social media outlets can easily identify and label these videos and images as such? - Should Companies be mandated to explicitly watermark every deep fake video to indicate it is a fake and not be taken seriously? - Should there be a digital fingerprint on every deep fake video produced to trace who produced such a video (for accountability and deter misuse)Please do pass on these questions for consideration by the Panel.

Pierre-Jean	VERRANDO	Eurosmart	On the 25th April 2018, the Commission defined in its Communication an European approach for Artificial Intelligence.This Communication sets out a European initiative based on a triple approach:1. Boost the EU technological and industrial AI uptake across the economy through investments in research and innovation and better access to data;2. Prepare for socio-economic changes;3. Ensure an appropriate ethical and legal framework.The Draft AI Ethics Guidelines of the High-Level Expert Group on Artificial Intelligence (AI HLEG) contributes to raise the awareness on the close relationship between ethics and technological choices in the digital age. When it comes to the development and implementation of AI, the deep interrelation between ethics and technology modify the way we usually think about technological advances by bringing together deferent disciplines: Ethics, Law, Technology, Industry and Cybersecurity.Eurosmart welcomes the European Commission initiative and the creation of the High-Level Expert Group on Artificial Intelligence. This initiative plays a key role when defining a common understanding of what the challenges brought by AI are. Organisations, value chain, their related threats and opportunities will be impressively impacted, AI's incidence on the cyber-resilience of our continent must be conscientiously analysed. AI is also challenging both the values and the governance of the European Union.Definition of AI---Eurosmart supports the provided definition of what the Artificial Intelligence is. This first achievement of the High-level Expert Group on Artificial Intelligence is a milestone to define common rules to make citizens, governments and businesses benefit from trustworthy AI. "Artificial intelligence (AI) refers to systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take (according to pre-defined parameters) to achieve the given goal. AI systems can also be designed to learn to adapt their behavior by analysing how the environment is affected by their previous actions.As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge, representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems" Based on this definition, the European legislator and the industrial and scientific community must nurture an ambitious approach to develop reliable AI based on the European technical know-how and on our common values in reference to the Charter of Fundamental Rights of the European Union.Can AI be considered as a product placed on the market? ----The accompanying document to the Draft AI Ethics Guidelines entitled "A definition for AI" describes what AI is made of. (figure2). Even if the provided elements are a very crude oversimplification of the state of the art, it does have the merit of once highlighting several essential	Is AI a Dual use?----Eurosmart underlines that a technology cannot inherently be ethical. It the way the technological application will be developed and implemented which defines its ethical aspect. Considering that this technology could be used at both civilian and military levels, for peaceful and military aims, AI could be a Dual Use in the sense of the Wassenaar Arrangement. Deep competences and full mastery of the AI technology is very crucial for the digital sovereignty of our continent. Eurosmart enjoins the High-level Expert Group on Artificial Intelligence and the Commission to further analyse this issue. By the way, the dual use can be also seen from attacker for cyberattacks in the combination of human and computer as well as from the defender of cyberattacks, for example in industry and in governments. Two examples: Intrusion Detection Systems (IDS) with learning function in industry and Chabot's in public services. Protection of Personal Data is a major ethical aspect----Eurosmart supports the approach adopted by the AI HLEG which is underpinned by the European values and the Charter of Fundamental Rights. These common values have inspired all the data privacy and all the Digital single market legislation. We recommend working on a more comprehensive statement based on the article 8 "Protection of Personal Data". This article allows the citizen to benefit from its personal data as an inalienable freedom and places the respect of this rule of law under the protection of an independent authority. We recommend that both ethical and technical aspects should carefully be monitored and guaranteed by an independent and trustworthy Third party. It shall be a key principle while designing and placing on the market any AI solutions.	Standards---The document mentioned technical and non-technical methods to achieve Trustworthy AI. Standards are put forward to ensure that qualitative and trustworthy solutions are indicated to the consumers actors and governments. Due to the sensitive nature of AI, standards must be carefully handled. The European Union should not enshrine in law any "private" standards or unilaterally business-driven initiatives which could lead to an imbalanced power relationship. It must be considered that AI technologies will fast growing, such an approach would deter innovation. Eurosmart enjoins the AI HLEG to rely on European and International standards to support the AI take-off. European Standardisation Organisations' (ESOs) work should be recognised as the primary reference for a trustworthy AI development. Eurosmart recommends referring to the Mustistakeholder Platform (MSP) for Standardisation while developing priorities for AI in the Annual standardisation rolling-plan. Both AI HLEG and the European Commission must pay attention to international standards for AI which are under development (ISO/IEC WD 22989) and standards resulting from ongoing work in ISO/IEC JTC 1, SC 42 on Artificial intelligence, as suggested and highlighted by CEN-CENELEC in their response to this consultation.Data processing and anonymisation ----Anonymisation of data must be effective and of non-temporary nature. The anonymisation mechanisms should not be "deconstructed" by AI.Personal data should be strictly anonymised once they are merged into a large data set. This process should also apply to meta-data, since they are blended with traditional personal records. AI has the capability to de-anonymize the same information based on inferences from other devices. Therefore, voice recognition and facial recognition could potentially compromise anonymity in the public sphere. In this regard, the distinction between personal and non-personal data should be clearly define in the draft guideline and shall comply with the rules enacted in the regulation (EU) 2018/1807 on the free flow of non-personal data when it comes to anonymized and scrubbed data. The draft AI guideline should provide at least some insights to better understand how to handle data processing with such a requirement level.The European Data Protection Board (EDPB) should also issue a concrete contribution through Guidelines on AI compliance with the GDPR. Moreover, as foreseen by GDPR, certification schemes for IA should be prepared. It is deemed necessary for producers and importers of AI solutions in the European Union.	A Cybersecure approach is more than necessary----Assessing Trustworthy AI cannot dispense with the definition of security requirements. The guidelines mentioned mainly safety driven concepts and requirements, which is not enough to protect assets of AI solutions and devices. Cybersecurity is key to prevent from potential attacks and manage the protection of critical assets. AI cannot be assessed against safety concept whose targets of evaluation are static. Therefore, we strongly recommend penetration tests by Humans as a fundamental component for assessing AI, to verify and stabilize robustness of the most critical AI applications.Moreover, robustness of IA implies resilience as well as reliability and reproducibility. Eurosmart supports the promotion of a cyber-resilient network in the Union to guarantee security by design and a functional assessment for edge-computing devices.Third party certification-----The international and European standards mentioned in the second chapter can be used to performed 3rd party certification. The European Cybersecurity Certification framework should be mentioned as the primary reference to assess trustworthy AI, the European Commission shall make it a priority in the upcoming Union rolling work programme for cybersecurity certification scheme.	Eurosmart strongly supports AI-HLEG's big step forward to define a common understanding for trustworthy AI. This initiative paves the way to AI development in respect of the European values in terms of data protection, privacy and cybersecurity. Eurosmart highlights the need to mention and to recognise the work on ESOs for a real EU added-value in terms of AI standardisation. Based on these standards, a real effort shall be made to assess the upcoming AI solutions. The European Union is currently deploying trustworthy certification mechanism through the Cybersecurity Act and the GDPR and should rely on it.

			technologies which underlie AI.- Machine learning is composed by data and their processing.- Robotics is manly hardware oriented- Reasoning involved embedder softwares.From the industrial point of view and regarding the future market evolutions, Eurosmart wonders if AI could be considered as a product in the meaning of the EU Single Market related legislations. As a product, the 1985 Product liability Directive 85/374/EEC would apply.The benefit of this directive lies in its balanced approach between the free movement of goods within the Union, the protection of citizen's safety and the empowerment of the economic actors. For a given product, the full liability is placed on the producer, the importer or the distributor of products. The same approach could apply to AI and thus, with the support of International and European standards.				
Anonymous	Anonymous	Anonymous	Totally agree that ethics has to play a foundamental role in leading AI development and regulation	I haven't seen the assumption that machine doesn't have to arm humans. Is it out of the scope of the guidelines? Important the ethics of the choices machines will have to do: in some situation better priviledge humans or machines? or which human should the machine prefer? (ie: the passenger or the pedestrian?) AI has to take and active step be Inclusive, not only Not discriminatory. Guarantee also the variety of knowledge. Not only what is in line with my preferences but also what is there "in the world". Not to limit the knowledge	Very important avoiding asimetry of information Very important well understanding and disclosure how the data are collected and how the data are interpreted and linked in order to get the "knowledge" and to veify it is not biased. Also very important to understand how context influence the data as source, and also how context influence the reliability of the output. Essential: education of the society, joung generation, institutions about AI and its underling features. Goverments should be committed to update their educational programs to spread instruments to people to allow them to deal with AI softwares. Should people be paid for the data they provide to the AI systems? In my opinion LAWS should be banned worldwide. Artificial Consciousness should be very undercontrol, therefore signaled when actions are going these ways Traceability is very important Inclusion, more than Non Discrimination Explaination about how data are structured and linked together Collection of data linked to "emotional" situations, more controlled and attentive Human Autonomy always guaranteed Give evidence of all "knowledge" not only what is in line with my preferences, in order also to foster the diversity of knowledge and its development Including the abilities of all the individuals (ie authistic individuals, different income or culture individuals, ecc)	Inclusion, more than Non Discrimination Explaination about how data are structured and linked together Collection of data linked to "emotional" situations, more controlled and attentive Give evidence of all "knowledge" not only what is in line with my preferences, in order also to foster the diversity of knowledge and its variety. Pay great attention to the coherence of data and how the knowledge is extracted by them, "Stop button" for self learing AI approaches higlhy recommended Implement very good systems for dealing with mistakes Fall-back plan: very important	I think that it should be tried to be avoid the oligopoly of data and application of AI techniques by few companies or institutions. In order to facilitate the awareness of the mechanism used bt AI system it could be imposed some disclosure of some parts of the algorithms, i order to make "consumers" more aware of how the "answers" they receive are deriving. The present application of Privacy is useless, we all accept whatever Terms and conditions in order to receive the service we want. Therefore for AI the regulation should be in the processes itself, in the people involved and in the literacy of the public about the problems and the hidden mechanisms. Which entity could control the right application of the Guidelines? Would it be possible to have a world wide discussion about general Guidelines applied world wide? Like for example a Davos meeting about AI regolamentation

Jacek	Ruminski	Gdansk University of Technology, AI Bay club	no comments	no comments	Respect for Privacy Digital records of human activities can be used to extract information that is not obvious for potential users. For example, from video sequence of face a pulse rate could be extracted and other vital signs can be estimated. It is obvious that users should be informed about the desired use of data processing. However, the potential risk of other data use is still very high. Data owners (or organizations that process data) can use data to build models e.g., to estimate how rich is a person (e.g., from cloths), how healthy is a person (e.g., biological vs. chronological age estimation), etc. Resilience to Attack. It is important to focus also on integrity (and security) of models (e.g., ANN/DNN models). This is very important, since data models can be easily exchanged, methods like "transfer learning" can be used, so there is a real vulnerability to attack – changing a model (e.g., that will not properly diagnose a person). Technical methods are also required to provide model's integrity and related control.	Technical methods are also required to provide model's integrity and related control. 8. Robustness: Resilience to Attack: What systems are in place to ensure model's security and integrity?	
Manfred	Bürger	ret. head Nucl. Reactor Safety Dept., IKE, Uni Stuttgart, now with InkriT (Inst. Krit. Theorie) and FST(Forum Soz. Technikgest.)	"The goal of AI ethics is to identify how AI can advance or raise concerns to the good life of individuals, whether this be in terms of quality of life, mental autonomy or freedom to live in a democratic society." (p.2). This is to be understood as checking where and how AI can improve human perspectives before just applying it, even before major development beyond research. Questions should be: Where are major problems? How can AI help?Further: "A domain-specific ethics code … can never function as a substitute for ethical reasoning itself, which must always remain sensitive to contextual and implementational details that cannot be captured in general Guidelines." This addresses a culture of reasoning, awareness and consideration which is capable to decide on applications and use for humans and society in specific cases. Thus, a main emphasis must be on developing such attitudes and thinking, such a culture, beyond just developing a set of rules. The emphasis must be on education and practice, practical learning in societal decision processes. Initiatives for this must be promoted by the expert group and the EC concerning major areas of societal concern, not just oriented at AI, but including possible use. Thus, this understanding of Trustworthy AI, also promoted by the document itself, goes beyond the framework of "ethical purpose" and "technical robustness" highlighted in the document.	A rights-based approach to AI ethics is certainly an important basis. But it does not avoid the need of developing a culture of democratic decisions about where to go with society and in specific areas. Specifically, it will not be sufficient to control self-learning AI, i.e. to solve the conflict with wished capabilities of autonomous action and restricting them to assure use for human well-being. This conflict can finally only be handled adequately by permanent consideration and controversial treatment in discussion processes of related working teams and in society. The risk beyond assurance by rights is that technical systems are increasingly trusted to guide vital processes and decisions, that the background of decisions and alternatives are no longer considered. And this, even as the basis of their decisions and learning processes remain unclear (black box) – adding hence one more source of undesirable developments. Attempts to retrospectively explore their "reasoning" (forensics) remain still inadequate. Therefore, and despite partial successes, the necessary role of humans must be emphasized, and the increasing transfer of human tasks to AI must be reviewed critically case by case, and according to needs and possibilities. This is a task beyond fixed rights and ethical rules. In particular, the necessary control and intervention levels and processes must be designed, and this design process should reach up to the level of what is societally wanted, to the aims of societal production and design, to the elaboration of societal well-being. This line may also be addressed by the formulation in the document (p.5):"Informed consent requires that individuals are given enough information to make an educated decision as to whether or not they will develop, use, or invest in an AI system at experimental or commercial stages (i.e. by ensuring that people are given the opportunity to consent to products or services, they can make choices about their lives and thus their value as humans is protected)."Freedom of the	The list of requirements for AI systems certainly supports concrete checking of specific AI devices. It may also help in deciding about use of AI devices in specific contexts to support human interests and perspectives. However, primarily, decisions must be based on needs and options to solve human and societal problems and to realize improvements. This should be the primary basis of trustworthiness (see also above). Only in the second step, the criteria addressed by the requirements should come into play. The use of the requirements should not replace primary choices by dealing only with additional assurance or justification of already pre-determined AI use (as with ethical considerations in the frame of unquestioned autonomous driving).Even to adequately treat the requirements within given frames, cooperative work is required, e.g. to adequately manage the data for self-learning systems which also requires feedback with control of results. This is also indicated under 4. concerning conflicts between user opinion and recommendation by AI, or under 5. with treating inherent bias in data (requiring awareness training addresses the cultural and cooperative questions).Robustness can also not be guaranteed based on design as also indicated under 8. by considering the whole product chain. This also means that processes of control and adjustment, cooperative processes are to be established.Safety also needs a culture of awareness. It cannot only be assured by technical design. Control must be permanent and under controversial views.Transparency (Explainability) can - especially with self-learning systems - also not just be reached by technical means. Since (or if) self-understanding and model-based understanding is not implemented but learning is just based on statistical pattern recognition, transparency can only be reached in rather limited way. This is also indicated in the section Testing & Validating (p.20), esp. by requiring that testing "should	The assessment list appears especially to be applicable for an AI device already chosen in principle, thus not addressing the questions and means to be considered before such a choice. Specific qualities of the AI system must of course be assured and the list supports this. However, the list excludes decision questions about the possible benefit of AI, where to use and for what. It addresses only good functioning of AI devices in a sense not acting against human orientation. This is only a narrowed concept of assessing trustworthiness as I further outline I my other parts of comments. A somewhat more special reference to the list and the questions concerning the envisaged use cases at the end of the chapter are given in my general comments.	The document gives general orientations for a human-centric approach to AI, including general guidelines on the realization of Trustworthy AI (concerning ethical purposes as well as technical robustness). Benefits and risks are specially addressed. This general view is to be supported, especially the emphasis to ensure that AI is human-centric and that "a continuous process of identifying requirements, evaluating solutions and ensuring improved outcomes throughout the entire lifecycle of the AI system" is required.However, this needs to be concretized for specific problem fields in order to be effective. Further, the use of the specific technic should not be set as given before controlling the risks and benefits. Instead, the choice and design of technical systems should be done according to the societal goals to be elaborated in societal discussion processes. This would be the democratic perspective to be combined with development of consciousness about the technical and societal development, including knowledge of alternatives. E.g., setting autonomous vehicles as given perspective narrows the systemic solution perspective of traffic, similarly with health and care systems as well as education (with given e-learning concepts). Well-being goals are narrowed by dominance of growth goals. Benefits and risks can then only be treated in such narrowed paths. Since a major emphasis in high-technology development including AI is on systemic treatment and organization, systemic solutions involving technics but primarily societal perspectives should be in the foreground. Good systemic solutions could even yield economic competition advantages compared to single paths of technical development (see systemic organization of traffic rather than just emphasis on autonomous cars and e-mobility, the latter even without considering means to provide electricity, especially if going into mass use, as well as raw material sources, e.g. for batteries).Self-learning systems need specific consideration, although the system approach itself already

individual as considered in 3.2 can in the AI context only be realized in ways indicated above, the same is valid for avoiding inference of AI systems with democratic processes (3.3). Not only "a right to a human-centric appeal, review and/or scrutiny of decisions made by AI systems" must be guaranteed and fixed or implemented in AI, but such processes must be organized. Internal and external experts as proposed under 4. (p.8) can support but do not replace an organization of cooperative work of teams really involved. External control cannot replace internal systemic control of those directly engaged in the subject (according to E.Ch. Wittmann, the only reasonable way to regulate a system is to reinforce its self-regulation). This is a typical feature of complex processes.P. 9: "AI can be a tool to bring more good into the world and/or to help with the world's greatest challenges." I agree in principle but this has to be elaborated case by case in processes considered above, involving decision processes in the whole society. Also: "AI systems should be developed and implemented in a way that protects societies from ideological polarization and algorithmic determinism" can only be reached by permanent processes considered above, not just by design. Environmental awareness requires more than control of given AI systems, but choice and design of systems, of society according such goals, as indicated in footnote 12.Footnote 13 as only direct reference to the working process addresses the requirement to "decide on how AI systems operate". This may be taken as support for my above argumentation but needs further elaboration and emphasis.For me, the critical concerns under 5. are not of major concern. They address needs, chances and risks which may only partly just be managed by rules without giving up chances or reinforcing risks. The evaluation depends to a significant degree on the status of society, on the degree of trust in the societal processes, therefore their good orientation and functioning. Therefore, I am stressing the establishment of a respective culture. Even the weapons question cannot be separated from the evaluation of risks. A solution perspective can only be to put emphasis on civil solving strategies of conflicts. Again, the problem cannot be reduced to that of an ethical design of AI. Human responsibility should of course always be maintained, but mainly in solving conflicts. If running into self-escalating conflicts, it may be too late and sheltering requirements may dominate and even require rapid, i.e. automatic and autonomous reactions and preparations for this. Solutions to avoid uncontrollable arms race have to be envisaged before.Also concerning 5.5, the major point is societal organization to keep human control. It can be assumed that for a long time there are still going to exist significant differences between human intelligence and "artificial intelligence", despite the increasing introduction of decision-making and even adaptive learning systems. The development of consciousness (self-reflecting capabilities as real intelligence) would be a stage at a quite higher level than the present implementations of learning behavior. The questions which really concerns presently and in predictable future is the progressive replacement of partial human capabilities by be performed by an as diverse a group of people as possible". It is also indicated under Traceability & Auditability as well as Explanation (p.21). It is to be strongly underlined and agreed that non-technical methods are also considered as necessary at all process levels and on an on-going basis, in addition to the technical ones. However, the non-technical methods are in the document mainly restricted to formal control means, rather than considering means to establish the necessary organization of societal processes and a respective culture. Introduction of representatives in charge of control is not sufficient. The point Education and awareness to foster an ethical mind-set (p.22) goes beyond this by requiring general informed participation ("to make people aware that they can participate in shaping the societal development"), also the following points indicate a wider view. But all this needs further reinforcement and elaboration – I consider this more important than adding additional points as asked for. Some of my essentials appear to be promoted in the Key Guidance block at the end of this chapter, esp.:• Make Trustworthy AI part of the organisation's culture, and provide information to stakeholders on how Trustworthy AI is implemented into the design and use of AI systems. • Ensure participation and inclusion of stakeholders in the design and development of the AI system. Moreover, ensure diversity when setting up the teams developing, implementing and testing the product. • Foresee training and education, and ensure that managers, developers, users and employers are aware of, and trained in, Trustworthy AI.

defines tasks of continuous supervision, of continuous design measures and risk management. Dealing with complex systems needs educated teams and a culture of discussion and consideration of controversial views (as can be learned from experiences with nuclear reactor systems). It needs various experiences, not only theoretical approaches. Autonomous actions of technical systems and human control must be permanently considered, evaluated and revised. This is especially valid for self-learning systems with their black box behavior. Methods to reveal the internal processes and to control them mainly technically, for example by setting limits, will not be sufficient. From this, a major area of concern - not treated in the document - is the design and organization of working processes. Educated teams with cooperative orientation, not restricted in their tasks by hierarchical dominance and including various views, are required. The shift in work to steering and controlling, programming work as shift to intellectual work, is not to be taken as a shift to a theoretical elite. Theoretically clarified experiences as basis to elaborate the essentials in complex processes are required. This elaboration can only be done in combined teams which must be sufficiently large to guarantee various views as well as elaboration under controversial considerations and discussions.Although already by this, work reduction should be limited, the tendency to job losses and technically produced unemployment, will remain, in view of the replacement processes promoted by technology, even going and even in reinforced way into the range of preparation and intellectual works, in general services work. These problems of general importance for the society are not treated in the document. A solution perspective should not only be seen in compensations by basic income concepts but should go into the perspective of general participation in processes to design society within the various problem fields to be also specified in the document. Education and development of consciousness and awareness of goals and problems must become a major goal. This is also essential for democratic perspectives in contrast to determination by technical restrictions and practical constraints, decided by experts. Means to go in this direction are to be considered in continued elaborations of the document.Trustworthiness of AI cannot be taken as a major goal if only selectively addressing the specific AI design. It must include the decisions about development and use, about contexts of use and kind of use wished. It must include the processes of dealing with AI as outlined here and also in general formulations indicated in the document. A societal culture of dealing with AI and technological fields of possibilities (fields of enabling) as well as risks has to be developed. Otherwise, ethical and human-centric orientation cannot be reached. Concerning the invitation on the last page of the document, to share thoughts on the assessment list for the 4 particular use cases of AI envisaged, it is again to be emphasized that the elaboration should not be limited to a list treating features of a given AI system but must be based on considerations about the basic organization of the specific areas and systems and the question how AI could be used with benefit to support basic human

digital systems and AI, already by the systemic approaches and even more by self-learning AI devices. This poses the question of human role in all areas of working and society and cannot just be answered in general by keeping human-centric emphasis, by keeping this human role in all processes. Even decisions are more and more delegated to machines. Thus, the human role must not only be stated but defined in general and for all specific processes in work and society organization. This definition has to be referred to the specific capabilities of human consciousness to develop models of the world, about human orientation and society, which has to do with understanding essentials even of complex systems. These capabilities are required to steer and control such systems with which we are increasingly concerned, by our own constructions and by increasing interference of them with nature. Such modeling capabilities are still quite different from the mere pattern recognition procedures of self-learning AI, although this may to some degree also be considered as model building. But, the major point here is understanding combined with the capability to derive essential features and to conclude on extrapolation possibilities of models as well as their limits. The present attempts with self-learning systems, even if quite successful, do not reach such possibilities. Their application - even attempted to physical problems - remains limited and doubtful with this respect. They cannot replace understanding based on modeling and derivation of essential features from complexity. Not astonishingly, failures of machine learning, of mere pattern recognition strategies refer to missing understanding, as e.g. in picture recognition revealed errors of bird recognition based only on items of the surrounding or the case of confusing a school bus with an ostrich due to small value changes as mentioned in the document (p.21). Learning based on increased data sets may help but only on this given basis and hardly concerning distinguishing different model areas, thus limits and extrapolation possibilities.Differences in the learning behavior of infants and animals are visible concerning human’s ability to transfer experiences and perceived patterns to other areas of experience - a knowledge that is acquired fast and already with a small quantity of data. This difference is true also with regard to the learning behavior of programs. This has probably to do with the human capacity to be able to find the essence behind the surface of things, the capacity to go beyond mere pattern recognition, leading ultimately, through the emergence of conscious modeling, to the ability to develop and reconcile complex thought scenarios in communication. Communication has been revealed as a key part in addition to capabilities of model development (understanding), both founding the differences between infants and animals, of humans and present types of AI. Thus, given that “AI” systems will not raise to such a level of intelligence, the danger lies precisely in the fact that such systems (with their current, still rather limited learning capabilities) and already less developed technical systems (without essential learning capabilities) are increasingly trusted to guide vital processes and decisions. And this happens, even though the basis of their

aims. E.g., AI in healthcare should not just be considered concerning robustness, reliability etc., but the basic question should be where it could support human aims and where direct human relationship should be kept. E.g., autonomous driving should not be primary, but the organization of traffic as a whole. With mainly public traffic organization, less private cars, less crowded streets, it would - by the way - also be much easier technologically to develop and use related partial automatic driving.In the assessment list, points 3, 4, partly 10 (purpose) appear to go at most in the indicated direction (7-10 are more important for final decisions on the choice of the specific AI system, also 1,2,5,6). Nevertheless, pre-phases of considerations and discussions about general intentions should be envisaged before just choosing and establishing an AI system. Some of the assessment questions may support this process, but do not replace processes of thinking about goals and alternatives in the specific areas. In view of the important specific areas of concern, esp. work processes, but also the general societal questions, I am astonished not to see representatives of trade unions in the panel of experts. I also miss institutions directly concerned with societal and working process questions as e.g. SOFI and ISF.

				decisions and learning processes remain unclear (black box) – adding hence one more source of undesirable developments.The human role in the future processes, more and more determined by machines, must therefore be defined based on the human capabilities to draw essential features from the complex processes, to understand based on modeling of experiences. These capabilities can only be effective in communication, in joint elaboration of essentials, based on experience and theory. Thus, establishing cooperative, collaborative processes is the key task, of processes which combine various experiences and views, various approaches, and yield evaluations and decisions based on controversial elaboration. Just in view of the successive replacement of human actions by machines/AI, it is necessary to establish control about decisions and processes, even concerning just avoiding failures and related risks. The human role must be established by emphasis on collaborative, cooperative organization and development of a culture of cooperation including the dealing with controversial views, a culture of elaborating essentials of processes and goals, a culture of human deciding at work and in society, a democratic culture. This would be the real basis for trustworthy use of AI and should also be the basis for possibly arising long-term concerns, e.g. with even more developed AI. Just considering such long-term questions may fail to solve the present and foreseeable problems with replacement of human capabilities and to keep and found human perspective.		
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential
Anonymous	Anonymous	Anonymous	Great vision!				Being a European, it makes me proud seeing how you / we stand up for our values by doing our best to make sure AI will be for us rather than the other way around.
Alla	Kos	Independent Consultant/Responsible AI	Positive: - clear focus on implementation, operationalisation and practical use of the guidelines as a living document - mention of the UN Sustainable Development Goals (SDGs) What could be improved: -Much stronger emphasis on the linkages to the SDGs needed, including reference to Agenda 2030 and the SDGs framework (key global/national framework for the #AIforGood)	Positive: -rights-based approach to AI Ethics -mention of importance of the rights-based approach to investors -recommended presence of an internal and external ethical expert What could be improved: -encourage investors to integrate environmental and human rights due diligence into their investment processes and decision making for ensuring responsible innovation -internal and external ethics *and* law expert (to emphasize ethics + law/human rights approach to AI Ethics and not reduce it only to ethics) - more decisively and explicitly use terms 'data protection and privacy' in relation to treatment of data	- One of the most obvious and highly recommended methods for realising trustworthy AI is the Human Rights Due Diligence (HRDD) process, specified in the UN Guiding Principles for Business and Human Rights, implementing the United Nations 'Protect, Respect and Remedy' framework. 1) authoritative global standard for preventing and addressing the risk of adverse human rights impacts 2) relevant to companies of any size, type of property, geography 3) HRDD is an ongoing process and in combination with the strategic foresight approach could be helpful in regularly assessing potential long-term concerns and dealing with uncertainty 4) HRDD includes remedy and in general serves as a risk management tool with focus on preventing and addressing harm/risks to people vs risks to business offered by traditional risk management approaches.	Again, assessment phase of the Human Rights Due Diligence process could be very helpful. It is in alignment with the process described in this chapter.	Good job on the first draft. Thank you for your dedicated and hard work!

					Engaging with the UN Guiding Principles on Business and Human Rights and the SDGs is one of the most meaningful ways for companies to contribute to the sustainable development on global/national/local levels.		
Peter	Rijnbeek	Erasmus MC					For trustworthy AI in healthcare extensive external validation is a prerequisite in my view. This requires a fully transparent and reproducible framework that enables development and validation at an unprecedented scale on our European data. Currently, this is not a reality because of interoperability issues of healthcare data and this issue needs to be addressed first. I think the external validity needs more attention in the guidelines.Furthermore, to avoid misuse and misinterpretation of predictive models we need to enforce minimum reporting requirements, that include for example fully transparent and reproducible definitions of the target population and outcome, modeling details, and a minimum set of performance measures. It is unethical to apply a model if these requirements are not met. The model could be applied to a different target population in which the performance is far from optimal. I suggest to give this point more attention in the guidelines.I support the focus on training of stakeholders in what AI is, and more importantly what it is not. Many people make causal assumptions on predictive models and we need to better educate all our stakeholders to avoid this. Personally, I do not like the term AI since it suggests something that it is not: intelligent. This is one of the reasons our field has a marketing problem and we always need defend its high potential to those who are not experts. Education will help to alleviate this problem and will make the future of AI in our big data era a gamechanger.
Stefan	Hügel	Forum Informatiker Innen für Frieden und gesellschaftliche Veantwortung e.V.	FIfF e.V. welcomes the High-Level Expert Group's draft on Ethics Guidelines for Trustworthy AI with its orientation towards the common good and fundamental rights, principles and values. We would like to call attention to some aspects we consider of importance and which should be stressed more. We share the High-Level Expert Group's concerns in terms of possible risks and consider an impact assessment and risk analysis necessary for those AI-systems impacting individuals, groups or the society as a whole. The interests of those developing, deploying or using AI dominate the draft document, whereas we think affected stakeholders should be considered in their role as groups and individuals impacted by AI. We suggest that impacted groups and individuals should be considered more prominently in the guidelines and especially in the use cases to be developed. What seems to be lacking in the Guidelines is an early warning mechanism about AI impacting social coherence. Even now platforms use any means to capture users' attention at all costs thereby leading to polarisation, radicalisation and	2. From Fundamental rights to Principles and Values: We consider the individual's "right to be left alone" a fundamental right. This is best implemented by design as the possibility to opt-in. It is to be preferred whenever possible. However, opt-out is the only possibility we found in the draft Guidelines. 3.5. Citizens rights: Collection and processing of behavioural data must be based on an informed consent of the individual. In order to systematically be offered express opt-out, governments must provide access to alternative processes. The Principle of Justice: "Be Fair": We support the requirement that AI systems must provide users with effective redress if harm occurs, or effective remedy. Regulatory provisions as to liability must be implemented by law, enforced effectively, and breaches must be strongly sanctioned. 5.1 Identification without Consent: We strongly object to face recognition or other involuntary and covert methods of identification using biometric data. They must be accompanied by strong rights of those impacted to be notified of the covert measure before being subjected to it. 5.3 Normative & Mass Citizen Scoring	2. Data Governance: We support the recommendations concerning data set tests and validation. 3. Design for all: We also support the recommendations concerning the accessibility and usability of technologies by anyone at any place and at any time, ensuring their inclusion in any living context. Footnote 24: Responsibility is defined only for levels (1) and (2), i.e. the developer. As for level (3), responsibility lies with design modelers, developers, data analysts, and for levels (4) and (5) with design modelers, data analysts, developers and human decision makers. 5. Non-Discrimination: Algorithmic price discrimination based on scoring is a current market practice that must be pushed back. 6. Respect for (& Enhancement of) Human Autonomy: Above and beyond providing explicit support to the user to promote her/his own preferences, systems must advise users that "nudging" is taking place and inform the users about the objectives to be achieved by it. 7. Respect for Privacy: Where personalisation takes place, it implies violations of privacy. Users must be provided with effective ways to turn off	Impact assessments and risk analyses must precede an ongoing evaluation in order to make it possible at all. This should not be left to stakeholders' discretion but be mandatory. 6. Respect for Privacy: Consumer convenience-orientated technologies are especially prone to function creep e.g. for surveillance purposes. Is there a mechanism in place to assure purpose limitation? 7. Respect for (& Enhancement of) Human Autonomy: Users must be informed about the objectives to be achieved by nudging.	

			fragmentation in the public sphere. This raises critical concerns and calls for impact assessments. We suggest that non-profit actors such as the Free and open software community be supported and incentivised to create AI tools for the benefit of citizens and independent groups to do assessments. We also suggest widespread information about the rights of groups and individuals impacted and a mechanism to ensure their informed consent in addition to defining. This may be accomplished by an obligation to label AI-applications as such and should be implemented ‘by design’. While compliance with applicable regulations is required, the Guidelines do not state the applicable norms, with the exception of the GDPR. For greater transparency and to facilitate discussion in the civil society, however, this would be necessary, especially since the Guidelines are not legally binding and not complying does not impose sanctions. Therefore, references to legislation are more than ever important with a view to international and transnational (co-)legislation, such as trade agreements. The Guidelines appeal to those developing, deploying or using AI to voluntary follow the Guidelines, establish codes of conduct etc. This is not a practice IT-monopolies usually abide by. We do not consider fundamental rights, transparency and the avoidance of harm as solely desirable values. They must be enforced effectively, and breaches strongly sanctioned to prevent harm before it is done.	without consent: We strongly object to this type of data gathering, analytics, and handling as it is a clear violation of peoples’ fundamental rights and may present security risks and result in breaches. Whenever citizen scoring is applied in a limited social domain, a fully transparent procedure must be available to citizens, providing them with information on the process, purpose and methodology of the scoring, and the possibility to opt-out of the scoring mechanism. Generally, an opt-in mechanism by design is preferable. The Principle of Explicability: “Operate transparently”: We welcome the requirements of an IT audit of the algorithm as well as a procedural audit of the data supply chain. 5.4 Lethal Autonomous Weapon Systems (LAWS): We strongly object to LAWS and support the European Parliament’s urgent call for the development of a common legally binding position.	personalisation and the accompanying collection of personal data. 1. Technical methods: The technical methods used should be incorporated in technical guidelines for particular domains, adherence controlled and violations sanctioned. In particular it is necessary to point out and differentiate the very methods used in learning systems, the methods for verification and testing and data used for the testbed. Traceability & Auditability: Training and empowerment for internal and external auditors must be provided and supervised, transparency is a key factor here. Explanation (XAI research): We welcome the topic in these Guidelines. However, we strongly object to deploying learning systems if no clear reasons for the interpretations and decisions of the system can be provided. Decisions may not be delegated to a system that cannot be fully explained. Regulation: We welcome the topic in these Guidelines. However, as mechanisms of liability and compensation are not in place as of now, we strongly object to deploying AI-systems in domains where human beings may be impacted until the time when they are.		
Dace Liga	Luters-Thümmel	European Women Lawyers Association	EWLA welcomes the opportunity to comment on the Draft Ethics Guidelines for Trustworthy AI. However, we regret that no stakeholder representing in particular the horizontal approach of gender equality aspects has been invited to work in the High-level Expert Group. Even though, the situation of business-to-consumer, business-to-business or public-to-citizen is mentioned and the need to a tailored approach to AI specificity a horizontal approach to gender equality aspects is missing in the paper (see Scope of the Guidelines, Draft p.3).	Having regard to the Charter of Fundamental Rights EWLA welcomes that the Draft emphasizes the necessary human-centric approach and the need for a consensual application of fundamental rights. However, theory and practice might differ considerably. In particular, we hold the view that the knowledge of (by vast majority) male developers of the specific needs or situation of women is extremely rudimentary (I.1. Draft p. 5). In addition, we would like to refer to I. 3.4 Equality, non-discrimination and solidarity including the rights of persons belonging minorities: even though one might assume that women are no minority, as well as consumers and workers are no minority, the wording leaves the impression that the issue and notion of equality has no gender related aspects. As already mentioned before, developers very often completely ignore women-related aspects. Some time ago, an incident at Amazon spread the word in the IT and AI community that their AI personnel selection tool had consequently eradicated all(!) female applicants from being eligible for employment even though this company already has a considerable number of female employees. It turned out that the data sets were outdated and more or less consisted only of male CVs. Another example for a biased view-point: The same could for instance apply to an invention of the automotive industry where for a long time only 'male' dummies were used to test car safety and security belts without taking pregnant women's bodies for such testing into account. We highly recommend to also	1. 1. Accountability (Draft p. 14): In case of a discrimination "an explanation and apology might be at least as important". No, not at all is this sufficient. Often the victims of discrimination even do not get to know that they have been discriminated by - for instance - being 'weeded out' by AI from the system like the Amazon personnel selection tool that threw out every female applicant. This is not just a matter of an apology, these are very real financial disadvantages on the labour market. The book written by Cathy O'Neil, "Weapons of Math Destruction" is a good description of possible effects of wrongfully 'feeded' AI. The same refers also to 2. Data Governance (Draft p. 14), it is a balanced data set that matters. There it certainly will be necessary to foresee a mechanism for corrections, as AI often does produce results, but those are not corrected afterwards. 5. Non-Discrimination (Draft p. 16): Here again the Draft mentions the direct or indirect discrimination of certain groups. Women are not just a "certain" group, but represent 50 % of mankind. 2. 1. Technical methods (Draft p.19): The notion of Ethics and rule of law by design is expressly welcomed. Compliance with ethical rules might mitigate the distortions that are created by biased data sets or biased programming.	III.5. Non-discrimination (Draft p. 25): a continuous testing for biases during development and usage of a system is highly recommended.	The European Women Lawyers Association strongly recommends to the High-level Expert Group to elaborate a gender-balanced concept for trustworthy AI.

				change the perspective and point-of-view on AI. The same applies to Chapter I.4. The Principle Do no harm (Draft p.9): not just children, minorities, disabled persons, elderly persons or immigrants are vulnerable, but also women can be vulnerable depending on their specific situation. A gender aspect is lacking in total. The same also refers to The Principle Be Fair (Draft p. 10). Developers and implementers need to respect not just minority groups and protect them from bias, but also discriminated gender groups. And again as already mentioned before: Longer-term concerns (Draft p. 12): as it is stated all current AI is domain-specific and requires well-trained human scientists and engineers to precisely specify targets. What if those are 98 % male? and lack the perspective of the opposite gender? There is a correction needed. The program is usually only capable to perform for which it has been programmed, in such a way biases are duplicated, if there is no corrective (human and opposite sex) instance.		As a general remark, this document covers measures pertaining to AI in the most advanced sense, i.e. artificial intelligence using deep learning, neural networks techniques that aim at making the AI system autonomous. We are using solely basic techniques like machine learning, the categorization of which under AI can be debated. The remarks / comments / answers below were made after careful reading of the document "DRAFT ETHICS GUIDELINES FOR TRUSTWORTHY AI".	
Anonymous	Anonymous	Anonymous	No particular remarks. As a general remark, this document covers measures pertaining to AI in the most advanced sense, i.e. artificial intelligence using deep learning, neural networks techniques that aim at making the AI system autonomous. We are using solely basic techniques like machine learning, the categorization of which under AI can be debated. The remarks / comments / answers below were made after careful reading of the document "DRAFT ETHICS GUIDELINES FOR TRUSTWORTHY AI".	The analysis was carried out by a group of several people with different profiles who have all an interest in the AI field. The comments are the result of a consensus of the members of the group. We believe that current regulations (RGPD) already provide important protection for fundamental human rights. We advocate prior information between a human and an AI because it is essential to know that we have an AI on the other side and not a human. We expect to attach the final guide to our future specifications that would contain AI, in this way, we will draw attention to the fundamental values of the EU which our public service must respect.	We believe that current regulations (RGPD) already provide important protection for fundamental human rights. We adhere to the 10 requirements. We do not have much experience in the field yet and therefore we can not give an opinion on the recommended methods to reach a trustworthy AI. The list of methods proposed seems to us a very good point of departure. Figure 3 is a good summary of the life cycle.	Within the scope of our projects, we believe that the questions in the evaluation are relevant and cover the whole field. Evaluation is a good tool for moving towards trustworthy AI. As a public service, ethics is at the heart of our approach. The crucial point for us has been that the system poses no risk to users be it individuals, businesses or public services. The product is an online recommender system. One of the risks of this type of product is to lock users into their past behaviors. In order to eliminate this risk, we have ensured that this kind of bubble filter phenomenon does not occur. For this reason, we systematically offer diversified content that does not come from the AI system in addition to the recommendations of the latter. The only consequence of an AI-related problem would be no more recommendations from him. Our system, a basic one, can't really be regarded as autonomous. The only part of autonomy they encompass is the machine learning part. In our system, the various risks are for the most part dealt with by the GDPR.	In the general opinion and for technicians, there is nothing shocking or that could lend negative feedback in the document.

Didier	Coeurnelle	Heales (healthy Life Extension Society)

The same is true for artificial intelligence as for many other technological advances. It can bring progress or it may bring human destruction.

However, when it comes to artificial intelligence it gets worse. In the worst-case scenario it is not only lives of men and women that are at stake, but the fate of humanity as a whole.

Prioritizing the use of artificial intelligence for everything that makes humans more resistant to disease and damage is a potentially positive endeavour for that objective and equally so as to reduce the risks of destruction.

Indeed, among the major characteristics of artificial intelligence, there is the fact that this intelligence is not necessarily:
= combined with a moral sense considering human life to be a fundamental value;
= endowed with 'common sense', that's to say, the same lines of thought of which most normally informed humans would spontaneously be capable.

For example : it is 'common sense' that the best way to look after plants so that they do not die of thirst is not to destroy the plants, even if this will avoid them dying of thirst. Yet, this common sense is not necessarily obvious to a 'super-intelligent' machine.

A use of artificial intelligence centered on health should reduce the probability of catastrophic consequences arising, since the ultimate goal will be the long-term improvement of human well-being. This will require us to theorize in detail all that is good for the health, resilience and integrity of human beings and this could also 'disaccustom' us, discourage us, from potentially harmful research.

Prof. Dr. Joachim	Fetzer	www.wirtschaftsethik.com	Using the European cultural area a specific AI developmentAI is a basic technology that will have a massive impact on coexistence and human civilization. Comparisons with the invention of printing are not exaggerated. A social learning process with AI is necessary. This makes it all the more important for European societies to bring their own traditions and values into the ongoing technological development and not to leave the shaping of change to others.The initiative for an "AI made in Europe" is therefore clearly to be welcomed in principle. The combination of technical innovations with values and traditions is important for the development of society and the market. Role of EthicsThe section "Role of Ethics" rightly describes different approaches, including teleological questions about the "good life" or deontological questions about a "good action". In the following chapters, however, an inappropriately shortened understanding of ethics is used, which is based on a rights approach. It is correctly emphasized that an ethics code can never be "a substitute for ethical reasoning itself". It should be added that liberal societies in particular are dependent on ethical reasoning and responsibility because they refrain from standardising the lives of citizens (and companies) too far and from legally controlling them. However, this line of argument is massively disturbed by the attempt at an assessment list in Chapter III (see Chapter III below). Trustworthy AI The approach with the "two components" (ethical purpose + technically robust and reliable) makes sense. Unfortunately, however, this approach is not adhered to in the following chapters (see the comments on Chapter 2).	The normative approach: The chart on page 6 is probably intended to signal that rights, principles and values are brought into an equally weighted context. This would make sense because it marks three different starting points of the ethics discussion, which only together adequately cover the spectrum of ethics. The text argues quite differently than the graphic suggests. The HLEG "believes in an approach" is already written at the beginning. In fact, there is a strong emphasis on the rights-based approach. From this rights-based approach, principles and values are then derived and, if necessary, subordinated to the rights-based approach. In the area of values, communitarian or even collective-utilitarian values are specifically emphasized. In particular, the European traditions of freedom remain underexposed. One would have to ask whether something completely different is not significant for Europe: The diversity of cultural value traditions, the mutual acceptance of which is a result of recent European history, is occasionally summarised under the heading "Unity in Diversity". If "Unity in Diversity" is specific to Europe, then this should be an important challenge for an "AI made in Europe". To what extent does social penetration with AI mean a "unification of behaviour and values"? This question is not easy to answer, but is not raised by the ethics guidelines. Instead, HLEG AI appears to the EU Commission as a standard-setter for European values that firstly do not exist in such a way, secondly are questionable in the composition presented here (an alternative approach would be, for example, "freedom - dignity - sustainability" as a triad of "European Business Ethics", Forum Wirtschaftsethik, Jahresschrift 2014), thirdly in combination with the methodological problem (see general remarks) can be dangerous (right up to unintentional totalitarian tendencies), and fourthly also very fundamentally determine how AI is dealt with.An approach that focuses on human dignity and freedom would set other priorities. The idea of freedom aims at the responsibility of individuals and organisations. Orientation towards the value of freedom must therefore mean making responsibility possible and promoting it. The broad use of AI can help here. But it can also be misused as an instrument for avoiding responsibility and self-empowerment. This is a relevant challenge. In the Guidelines (centrally e.g. page 5 above) freedom occurs only in a tamed form of "democratic freedom" and is quite insignificant overall. Chapter I.4, Principle "Do no harm": A "possibility to refuse AI services" that must always be enforced at every point must be rejected in the long run. The underlying problems are currently sloppy introduction processes of AI (self-criticism in the industry would be sensible) or a lack of user training. The existence of a personal complaints authority in monopoly situations (e.g. with authorities!!) must remain a fundamental right (due to errors never excluded). However, if competition is functioning, there is nothing to prevent a provider from selling tickets only via AI as long as there is another provider who also serves other customers (perhaps at a higher price). There is no right to "everything stays as it was".Section 1.4, Principle of Autonomy: "Preserve Human Agency"The principle of autonomy and	Before the points II.1 to II.10, something more fundamental is to be demanded, which only appears hidden in Chapter III. This Chapter II.0 could be: The use of AI must always take place in clear responsibility. This responsibility can be individual or corporate (parents are responsible for their children, companies are responsible for the results of the AI they use). The type of transparency, traceability, explainability or perhaps randomness required depends on the respective area of application. It is noticeable that from Chapter II onwards the distinction between the two aspects "ethical purpose" and "technically robust and reliable" no longer occurs. However, this would make sense. The aspects "technically robust and reliable", to which sections 8 (robustness) and 9 (security) are most likely to be assigned, should be expanded. AI is a specific tool that functions in interaction with people and groups or in social processes. Technical "errors" therefore also include interaction "errors". Therefore the point "transparency" should also be inserted here. Note to Chapter II.10 ("Transparency"): The last sentence in this section would be the end of any business secrets and own business models. For pretty much all models should be said to have "use human data or affect human beings or have other morally significant impact.	This section, which has been specially marked as "provisional", contains an important discussion point: Governing AI autonomy (4): Further diffusion and avoidance of responsibility is the greatest danger resulting from the use of AI systems. The intention of the demand "to ensure that an AI system always makes decisions that are under the overall responsibility of human beings?" is therefore correct. On the one hand: AI systems make no decisions. On the other hand: It is decisive that the responsibility for the results of algorithmic procedures can be attributed. But not only human beings can be considered for this. Responsibility can also lie with companies or other corporate actors.But overall, the draft of this chapter should be viewed with scepticism:(1) Too early: The attempt at such an assessment list comes too early in the current historical situation. The provisional character is indeed emphasized. Nevertheless, such lists have their own dynamics. The choice of the "rights approach" in Chapter II particularly suggests the danger of a legalistic misinterpretation. Whether this interpretation - contrary to proclamations of HLEG - is even intended would be speculation. (2) Too unspecific: Such assessment lists are quite useful in the context of sector- or even company-specific value- or integrity-management. As a one-size-fits-all instrument, on the other hand, they either seem hostile to innovation or remain superfluous and functionless. The former (hostile to innovation) would be bad for "AI made in Europe". Example: Does a "design for all" (3) make sense in the industrial B2B sector? The latter (functionless) would be a disservice to the serious approaches of working with ethical guidelines and value management.(3) No distinction between process and goal: It would be helpful to outline a social learning process and the tasks of developers and operators of algorithmic processes in this learning process, which also includes trial and error. It would be good to describe this process and outline guidelines for different actors in this learning process. This also applies to state institutions whose own handling of AI (keyword: eGovernment) could become the driver of a European AI. Only in the course of time will some aspects prove to be necessary and capable of regulation, which then have to be defined and enforced not in ethics but in law.	Methodological Problem: Ethics as a Discourse on Norms ("Ethics 1") and Ethics as a Mode of Control ("Ethics 2") There is no legal definition of ethics. But there are two understandings of „ethics‟:Ethics as norm discourse (Ethics 1) is the reflection, definition and content-related interpretation of the major concepts: Freedom, justice, human rights, the common good, the "good life", prosperity, etc. These terms represent an invitation to consensus and are therefore usually described only "thinly". They are and remain open for interpretation (What is meant by the common good?) and for different weightings (In doubt for freedom or for security?). It is desirable to keep on doing this. At the same time, such a canon of values is in constant movement and never fixed.Ethics 2 is a mode of self-control of society: more or less voluntary or by social pressure suggested consideration of moral values and moral consensus. Ethical behaviour is one that is based (more or less) on itself and aims at self-control. Ethics 2 competes with law. What is to be worked with which mode of control? What through regulation? What with incentives? What through market surveillance? What through criminal law? What must remain the subject of moral self-regulation, otherwise other values will be endangered. Ethics 1 never competes with law. For some aspects of the canon "Ethics 1", the law is an instrument for enforcement. For example, important aspects from the area of "Do no harm" are not left to voluntariness, but must be legally defined and enforced. For other aspects, especially for the category "Do good", however, the law is unsuitable. Ideally, particularly important aspects from an ethics canon are enshrined in law, mostly the negative ones: Not killing. Whoever tries to legally enforce "Doing good", on the other hand, creates a society in which personal freedoms are only possible in narrowly standardized channels.The draft of the Guidelines mixes this continuously and thus creates massive misunderstandings. In Chapter I..4, for example, "Do good" and "Doing no harm" stand side by side as two of 5 principles. Nowhere in the following up to the assessments in Chapter III will the different modes of implementation (market, law, incentives, moral pressure ...) be reflected upon. Because of this ambiguity, all proposals must be read as if they could sooner or later serve as the basis for general legislation. A legal enforcement of the ethics guidelines - as they are - would not be the beginning, but the end of every European AI. The HLEG repeatedly points out that this is not the intention. However, in the absence of a methodological distinction, this interpretation is suggested. This damages the credibility of the ethical guidelines.Understanding the societal learning process: "Trustworthy AI" or "Societal Trust in times of AI"?AI as a new basic technology means a learning process for all. In this process, different learning abilities, willingness to learn and responsibilities have to be discussed. This learning process includes everyone: Teachers and students, developers and users, businesses and consumers. In this process, companies have responsibilities that they are not always fulfilling. But the learning process also involves consumers. They must also be given the opportunity to do so. To this end, the disclosure of the use

				"Preserve Human Agency" is worth supporting - as is the shared responsibility of companies that use AIs for human autonomy. The way we talk about AI has a significant impact on respect for this principle. The Guidelines themselves should contribute to this. At least in this document, formulations such as "AI Decision Making" or "AI Decisions" should therefore be systematically eliminated and replaced by "AI Processes" - "AI Process Outcomes".Example: page 9 bottom/10 top: If one is a consumer or user ..... a right to decide to be subject to direct or indirect AI (replace "decision making" with "process-outcomes") .....Example: page 25, point 7, indent 3: Does the AI system indicate to users that a decision, content ... is the result of an algorithmic [replace "decision" by "process"] of any kind? Behind this proposal is a criticism of the definition of AI, as presented by the HLEG in December 2018 and in which AI is believed to have at least instrumental decisions. The Konrad-Adenauer-Foundation will soon publish a discussion paper on this subject: "Algorithms do not decide and will never do so". A short version: https://www.forum-wirtschaftsethik.de/mein-hund-die-kuenstliche-intelligenz-und-ich/ . Section I.5 "critical concerns": The main risk does not lie in the emergence of a strong AI. This can be left to philosophical speculation and science fiction. The greatest challenge is that society does not master the learning process or that the basic idea of responsibility is lost in the fundamental process of change. There are numerous indications of this. This has been explained and explained elsewhere. This aspect does not occur at all in the ethics guidelines.	of AI is important so that everyone in society can learn how to handle AI processes. A right to refuse interaction with AI, however, must be rejected - at least in competitive markets - (if there are suitable appeal bodies in case of conflict) and cannot be upheld anyway. There is no right to "everything remains as it is".In the draft Guidelines, the consumer appears to be almost always an object to be protected. This is probably due to the chosen "rights approach". The consumer does not appear to be the subject to be empowered. This achieves the opposite of what should be achieved with an ethical approach to AI. Artificial intelligences are a new important element of social interaction. So it is not only about inserting a "trustworthy AI" into an unchanged society, but also about promoting "Societal Trust in times of AI" and shaping the development, implementation and use of AI accordingly.Translated with www.DeepL.com/Translator
Mate	Szarka	Vitrolink		The general guilines and ideas are modern and human centric. It Should be driven this way! However logically I would suggest some restructuring. Not Ethics should come first, but data. Thrustworthy data should come first. The data should be available to the industy in a secured way and let it's members create from the same structured/checked data pool an ethical AI instance with ethical decision making and letting AI act only as a consultant. Humans should be hold accountable for the decisions their softares made from the legit data.

Hélène	Beauchemin	Stradigi AI			Identification without ConsentAs the guidelines rightly mention, consumers often provide their consent without consideration. In order to address this issue, we believe that in addition to developing efficient means to provide consent where possible, other means must be considered and implemented to protect consumers. The reality is that as technology becomes increasingly complex, it is difficult, and in many cases, impossible, for consumers to provide clear and valid consent to the use of their personal information. Companies therefore face seemingly incompatible standards: meeting high regulatory burdens to comply with personal information regulations, while providing the best customer experience. Lengthy privacy policies and click-through agreements are meeting neither of those objectives. In addition, recent research has demonstrated that re-personalizing "anonymous" data is relatively easy with today's technology. We therefore urge the Group to recommend that AI stakeholders consider additional means of protecting the privacy of consumers beyond standard legal documents, through means such as differential privacy and creative consent mechanisms. For example, differential privacy provides a "mathematically provable guarantee of privacy protection against a wide range of privacy attacks, i.e., attempts to learn private information specific to individuals from a data release. Privacy attacks include re-identification, record linkage, and differencing attacks, but may also include other attacks currently unknown or unforeseen." This means that using differential privacy, individuals are protected against re-identification through a mathematical screen created in the dataset, making it impossible to single out an individual in a dataset thereafter. That being said, differential privacy is not suitable for all instances of privacy risk and many factors such as privacy risk tolerance, dataset size, and exposure of employees to raw data should be taken into account. Such methods, combined with strong information security processes, clear and easy to understand consent mechanisms, will allow for AI to gain trust among consumers. Furthermore, it's vital to remember that different AI products may have different impacts on individuals (more on this topic at the end of this document), and as such, different ethical frameworks may apply. For information generally considered as more sensitive, such as biometric data, experience in other jurisdictions, such as the State of Illinois in the United States, informs us that determining where to draw the line is not an easy feat. Indeed, there is currently a split among the Illinois appellate courts regarding whether an individual has standing to sue following a company's mere technical violation of the Illinois Biometric Information Privacy Act ("BIPA"). The question of whether non-compliance with the obligations provided in BIPA highlights the difficulty in determining how we can achieve the integration of societal values into legislation. As such, the main question at issue in this case is whether an individual is sufficiently aggrieved automatically by the fact that a company has not complied with BIPA, from the moment biometric information is collected without the required statutory notice, or if actual harm must be alleged. A policy reflection is needed to determine how	We recommend considering the following Assessment questions when evaluating the specific use cases in question: Insurance Premiums: AccountabilityWhat formal processes (complaints and appeals) exist for users to dispute decisions made by the AI system?What formal processes have the company put in place to ensure that third-party data has been sourced appropriately and all users provided informed consent?Data GovernanceWhat measures have been taken to isolate different data sources from each other and minimize the risk of re-identifying individuals via data triangulation?What measures have been taking to ensure timely retraining so that predictions remain up to date? Governance of AI Autonomy (Human oversight)What measures exist to ensure that human operators will not interfere with the system, and what standards exist to allow employees to dispute decisions made by the AI system?We would delete the word "always" and specify the following in this question (in italics): What measures have been taken to ensure that an AI system makes decisions that are under the overall responsibility of human beings, as applicable and particularly when such systems have a substantial impact on individuals?Respect for PrivacyWhat measures have been taken to consider algorithmic guarantees against re-identification privacy such as differential privacy?What measures have been taken to ensure that researchers, data scientists, and developers have limited access to data with personally identifiable information, or where re-identification is possible? RobustnessHow is the company versioning and maintaining documentation on different models applied to different scenarios to ensure all provide a similar service experience?What assessments have been made to strike the appropriate trade-off between communicating model inputs to the public and ensuring users do not "game" the model? TransparencyWhat certifications, standards, or audits has the company undergone to demonstrate its adherence to Trustworthy AI?Healthcare Diagnosis and Treatment AccountabilityWhat measures have been taken by the device manufacturer to ensure medical users are fully informed of reasonable application boundaries?Have regulatory agencies sufficiently established bounds for bias within performance of the medical device?Who is accountable in the case of a erroneous medical decision, and what guidelines have been put in place to ensure proper operation of the medical device?Data GovernanceWhat measures have been taken to ensure that models are updated if users choose to revoke their consent? How should a company balance the right to privacy and data suppression with the need for a medical device to perform consistently?Has the organisation explored methods such as federated learning to allow users to maintain control of their data without making large sacrifices in model performance? Non-DiscriminationWhat measures have been taken to reduce bias in the training data? Which frameworks for fairness are being applied and why? Respect for (& Enhancement of) Human AutonomyAre there specific circumstances where humans are not allowed to overrule medical decisions made by the AI system? Respect for PrivacyAre there points of no returns, where if after a certain time period	Focus on Impacts One consideration that we urge the HLEG to take when drafting their upcoming policy recommendations is to shift their focus towards making an assessment of the real world impact that AI systems may have prior to the start of any formal and ethical review process. We would like to point the HLEG towards efforts currently in progress at the federal level in Canada to standardize the use of Algorithmic Impact Assessments for any and all AI systems that are deployed by the government to ensure appropriate governance. We believe that adopting this framework provides several unique advantages:Without a focus on impact, we are led to believe that, prior to a rigorous ethical assessment, all use cases are equally worth assessing and that their potential for harm is the same. We caution the HLEG that this may introduce unnecessary regulatory overhead, and that the legal and financial burden may be especially pronounced for smaller firms. We worry that if these guidelines were to be adapted into policy and mandated by law, companies may implement fewer AI initiatives, even though some may not pose significant risk of negative impact on any individuals in particular or to the public at large. We believe that it would be more effective if the adopted assessment framework recognized that different applications and use cases, such as the implementation of a product recommender on an ecommerce site and the rollout of full-scale autonomous vehicles, should be evaluated under different criteria, based on potential harm or risk of harm. Shifting the focus towards potential impacts allows for a standard assessment framework across all applications and uses cases without the need for individual Enterprise Ethics Review Committees to adopt new assessment criteria for each use case under consideration. This reduces the amount of work that is required of Enterprise Ethics Review Committees, which we believe will subsequently result in more frequent use. A standard model will also provide a common language for assessing different applications and use cases, and therefore allow for a better understanding of which applications are more or less appropriate based on previous implementations that resulted in either success or failure. Footnotes6 - "Algorithmic Impact Assessment (v0.2)" https://canada-ca.github.io/digital-playbook-guide-numerique/views-vues/automated-decision-automatise/en/algorithmic-impact-assessment.html (Accessed January 31, 2019)

best jurisdictions wish to address this question, based on the importance that privacy represents as a societal value. Covert AI Systems Although we broadly agree with the Group's position that a human should in principle always have the right to know whether they are interacting with a machine or an intelligent agent, we foresee many edge cases that may complicate the implementation of resulting policy recommendations. We advise the Group to look towards existing measures concerning information disclosure, under acts such as the forthcoming California Bot Law, SB-1001, which requires (primarily) social media bots to disclose their artificial nature. Legislators anticipate copious litigation to deliver resolutions to these edge cases, but we would like to present several to the Group for consideration that may not arise or are not directly pertinent to SB-1001. Should disclosures be enforced on all recommender systems that provide recommendations to people in domains such as e-tailers or insurance companies? If so, how exactly will AI Systems be defined, and which technologies will be encompassed under the definition? Is it appropriate to compel speech in instances where there is limited risk of harm? Although we admit that in instances like tobacco labelling it would be prudent to compel manufacturers to declare adverse health benefits, do most Covert AI Systems meet the same standard of harm? Should compelled disclosure only apply in situations where an affected individuals stands at risk for harm or damage? For example, bots are also used for all sorts of ordinary and protected speech activities, such as poetry, political speech, and satire: mandated disclosure would restrict and chill the speech of artists whose projects are based on having a bot performing these tasks. How should hybrid systems be dealt with, where although a human is executing the decision, a strong recommendation has been made by a Covert AI System that is almost always adhered to by the human operator. We present again the example of an insurance company, where the decision to offer a certain policy price is administered by an AI System, but the decision is executed and communicated by a human. Does the right to disclose hold in proximate cases such as this?In addition, by mandating disclosures in all contexts, this may put at risk the speakers who wish to remain anonymous, such as marginalized or traditionally silenced groups and individuals, a chance to be heard through a different medium like a bot. Finally, there are significant difficulties in enforcement. For example, as mentioned by some organizations, "platforms can try to use metadata like IP addresses, mouse pointer movement, or keystroke timing to guess, but bot operators can defeat those measures. These measures can also backfire against certain groups of users—such as people who use VPNs or Tor for privacy, who are often inappropriately blocked by sites today, or people with special accessibility needs who use speech to text input, whose speech may be mislabeled by a mouse or keyboard heuristic. Platforms can also try to administer various sorts of Turing tests, but those don't work against centaurs, and bots themselves are getting quite good at tricking their way through Turing tests." Taking a step back and projecting into the future, we

or level of involvement, users are no longer able to revoke their consent? What is the procedure for data suppression when models have been frozen, for example during regulatory review? RobustnessWhat measures have been taken to ensure robustness to different conditions such as blur, lens flares, different resolution images, and adversarial attacks?What measures have been taken to reduce the exposure that potential attackers could have to model weights, class probabilities, and prediction pipelines? TransparencyWhat measures have been taken to communicate the biases of the medical device to users for whom the device may underperform?Footnotes5 - Electronic Frontier Foundation, " Should AI Always Identify Itself? It's More Complicated Than You Might Think", https://www.eff.org/deeplinks/2018/05/should-ai-always-identify-itself-its-more-complicated-you-might-think (accessed on Jan. 31, 2019).

				also envision a world in which most, if not all, business decisions are likely to fall under the purview of AI Systems. Should at this point all systems continue to declare their artificial nature, or does this guideline merely serve as a bridge between the present and the future?Footnotes1 - Kobbi Nissim, Thomas Steinke et al., “Differential Privacy: A Primer for a Non-Technical Audience”, https://privacytools.seas.harvard.edu/files/privacytools/files/pedagogical-document-dp_new.pdf (February 14, 2018).2 - See Rosenbach v. SIx Flags Entertainment Corp., No. 2-17-0317 (Ill. App. Ct. Dec. 21, 2017) (in favour of allegation of actual harm to provide standing); Supreme Court reversed the appellate court decision and confirmed that no actual harm was needed to provide standing (2019 IL 123186 (Ill. Jan. 25, 2019)); Sekura v. Krishna Schaumberg Tan Inc., No. 2-18-0175 (plaintiff has standing to sue without allegation of actual harm).3 - Such obligations are: (1) notification of collection; (2) requirement to obtain consumer’s consent to such collection; and (3) requirement to have a written policy setting forth a retention schedule and guidelines for destruction of the information.4 - “Senate Bill No. 1001 CHAPTER 892” https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001 (accessed January 31, 2019)		
Anonymous	Anonymous	Anonymous	Building up trustworthy relationship means take into account opinions and views of those who are involved in such process. Thus, Guidelines should be discussed and approved by society at large and not only by stakeholders if the final aim is to create trust in the benefits of AI. A pan-European consultation on this should be taken into account.	In the Chapter, Oviedo Convention has been mentioned as clear example of EU grounding of ethical principles in biomedicine. But not all the state members have endorsed such principles, some have not ratified it and some (i.e. Italy) have not produce the internal Law for an effective application. Thus, it should be better not to not mention that example since really highlights that it is very difficult to identify EU common values in the governance of emerging technologies that can underlie such relevant guidelines, without instruments and means to verify what are the values at stake and what can be considered an EU value. Just structured and robust methods of consultations can provide a clear picture about this.			
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential

Christopher	Dimech	Naiad Informatics	If citizens don't control the technology they use, the software codes embedded in their computes and devices controls them. Public agencies exist for the people, not for themselves. When they do computing, they do it for the people. They have a duty to maintain full control over that computing so that they can assure it is done properly for the people. They must never allow control over the state's computing to fall into private hands.The document "Draft Ethics Guidelines for Trustworthy AI" composed by the European Commission's High-Level Expert Group on Artificial Intelligence was made public on 18 December 2018. We may be standing on the verge of a transformative strategy that focuses onpractices and innovations that empower people in terms of freedom, privacy and proper oversight; rather than guaranteeing unjust privileges to businesses, corporations, and governments when the universal inalienable principles of human rights become a threat to their profits or interests.Information and knowledge is power. One of the biggest exploitative casewhen it comes to this phenomenon has been the Cambridge AnalyticaDigital Data Leak Scandal} that emerged in March 2018. A subsidiary ofStrategic Communication Laboratories (SCL) Group, EuropeanCitizens had witnessed shareholders of the company being served withNational Order of Merits, Medals, and Lavish Meals. During the 2005 Defence and Security Equipment International Arms Fair (DSEI) - the world leading military technology and weaponry Trade Fair that attracts high level government ministers, senior military defence staff, and leading figures in the global warfareindustry - SCL Group demonstrated its capability to orchestrate and launch military disinformation campaigns, psychological warfare and influence operations through mass deception. Their case study focused on a dystopian and tyrannical mass surveillance and control system directed on the inhabitants of big cities like London.But the profusion of technology companies providing equipment andsystems to support the tyranny of Nation States around the globe extendsfar beyond the data leakage and surveillance capabilities of SCL.Italy's Area Spa, for instance provided Bashar Hafez al-Assad all therepressive technology he needed to drive his surveillance state inSyria. In another case, Mubarak's secret police intercepted internetnetwork layer protocols to spy on voice services such as Skype, Viberand WhatsApp used by citizens in Egypt. Following the overthrow of HosniMubarak, Egyptian resistance fighters discovered a license contractbetween the Arab Republic of Egypt and British-German manufacturer ofsurveillance and monitoring systems Gamma Group supplying maliciousbackdoor controller software that allows unauthorized access to affectedcomputers (see 1. https://www.zdnet.com/article/top-govt-spyware-company-hacked-gammas-finfisher-leaked/ 2.https://www.propublica.org/article/leaked-docs-show-spyware-used-to-snoop-on-u.s.-computers ; 3. https://www.f-secure.com/weblog/archives/00002114.html).Consequently, any talk of ethical policy with respect to ArtificalIntelligence at the European Commission Level must be based	One of the biggest challenges in the digital world is the ability for users to divorce themselves from physical identity rather than continuing with the current system of total surveillance. Particularly, surveillance imposed on citizens have reached levels that are incompatible with universal principles of human rights and behaviour, putting legislators at the European Union on the wrong track. For instance, taking interpretations from Joseph Cannataci, United Nations Special Rapporteur on the Right to Privacy, without asking those with more credibility to help the Union think things through is a fatal mistake. Cannataci is an observer whose legal insights and interpretations are useful, but computer hacker sources such as the Free Software Foundation and the Chaos Computer Club are better suited to illustrate the reality as they have direct experience in digital privacy and security. Cannataci has in the past criticized the use of facial recognition technology Yet, instead of realising that the United Kingdom has become an Extreme Surveillance and Security State, he still thinks that the Investigatory Powers Bill the UK legalised in 2016 can be made acceptable to civic society by writing up privacy and data protection clauses in a way that provide enough oversight mechanisms that protect citizens effectively.For instance, today, when people connect to the internet and visit websites they are being continually bombarded with Cookie Consent Notices and forced to agree to unjustly imposed Terms of Service. Once users consent, the digital cookies along with Javascriptcode instructions embedded in webpages track user activities in specificweb browser windows. Furthermore Persistent Cookies continue to identify, track, gather, and store the browsing activities of citizens across multiple websites and browser windows, even after restarting the browser. Thus, instead of the current EU's approach of regulating through its General Data Protection Regulation (GDPR), laws must be enacted that stops systems from collecting personal data or leaking digital information throughout the network. The concept of Society Under Surveillance enables citizens to mount a moral, ethical, social and political critique of the information processing practices in our society.For decades, abusive surveillance machines of proprietary companies. have watched people by utilising malicious functionalities in software, but also more recently in hardware. In particular, surveillance has spread dramatically away from the keyboard, in the mobile computingindustry, in the office, at home, in transportation systems, and in the classroom. For instance, the monitoring of communications is maintained in many universities through a massive surveillance system where the communication and technology businesses partner with governments, both foreign and domestic with their imposition of the overarching Google Accounts students and staff are required to use (seehttps://www.um.edu.mt/itservices/policies/jcstudent).If the union values the freedom and autonomy of all citizens, it cannot allow member states to set aside the rights of citizens, for instance, with the surveillance tactic of geolocating SIM cards or the installation of Face Recognition Technology and the (see 1.	There are several extremely serious problems associated with the lack of respect for human autonomy in the European Union. For instance, the European Parliament has on September 12, 2018, approved and passed the copyright legislation that is instrumental in continuing to shape a kind of society where citizens have no rights in the digital world. MEP Axel Voss, the head of the main Legal Affairs Committee (JURI), together with other loyal MEP's such as Francis Zammit Dimech, another member of the Legal Affairs Committee, continued his aggressive lobbying until they succeeded in influencing elected officials to vote in favour of the copyright directive. This brings to prominence the brutal attitudes of those who seek to serve the overall strategic and business interests of big corporations, big stars and big media conglomerates.Tracking people is the basis for the Surveillance and Control of Citizens, where digital filters are used as a type of Digital Restrictions Management system that restricts citizens by monitoringtheir activities. The new EU copyright directive forces platforms that accepts user-generated content to filter out stuff that might violate someone's copyright. It involves the blatant monitoring of what citizens read, how they read, and linking that information back to them,effectively creating a Union that mistreats its citizens. Tracking searches for books and keeping records of book purchases is also wrong. Books and EBooks are today taking away many of readers' traditional freedoms.In 2009, Amazon remotely erased Orwell Books "1984" and "Animal Farm" from Kindle devices of readers who had bought them. Is this not tyranny? Imagine if this happened in the physicalworld - A publisher who forcefully storms into people's houses in order to find and take their book. We then move on to the trend in copyright legislation to extend the duration of copy restrictions to effectively perpetual copyright terms that nullify the intended effectand violates the spirit of the "copy restrictions for a limited time period". The lobbying power of corporations have convinced governments that the national and regional patrimony of innovation and creativity is best protected in the service of their commercial interests. A view that misinterprets copyright, to benefit publishers using the unjust argument that if a certain practice is reducing their sales, or they think it might, we presume it diminishes the number ofpublications by some unknown amount, and therefore it should be prohibited. As a result, citizens have not been amused by the jubilations of MEP Axel Voss and MEP Roberta Metsola on the passing of the European Copyright Directive. Even more disconcerting have been the amendments by legisator Therese Comodini Cachia, who was instrumental in introducing the power of publishers to sue in the name of the authors. She betrayed all European citizens by imposing a restriction on all European citizens, controlled mainly by the publishers, in the name of the authors.The costs to society now outweigh the benefits of copyright. We are being led to the outrageous conclusion that the public good is measured by publishers' sales. Doctrines that bring into European Law changes to the implicit presumptions and nature of our most fundamental legalstructures, in order to control the ability of citizens to actively participate in a digital	Contemporary technology and democratic control have a difficult relationship. As hardware and software present in most of the technology we use (in game consoles, in smart phones and mobile computing, in tablet and personal computers) continue getting nastier, surveillance-infested products deserve contempt and disrespect (including the companies that produce them and their lackeys in politics and law), and ought to be made illegal. We must not be distracted by statements of what the state or companies will do with the information they collect (e.g., by policy clauses claiming that only aggregate, non-personally identifiable information is shared with third parties).The sharing of information was different before the advent of the networked technology. A couple of decades ago, one could distribute information by printing it on paper and handing it out. This was the way Samizdat, a form of dissident activity across the Eastern bloc, reproduced censored and underground publications. Information was disseminated by hand, from reader to reader. Today, if you wish to share the information on the internet, you need the cooperation of companies, such as Internet Service Providers (ISP), Domain Name Registrars and Hosting Companies. And they can cut you off arbitrarily without having to provide a reason and have your service termination justified in court.The spectrum of citizen rights is incomplete without establishing inalienable universal human rights in cyberspace as well. In practice, this means the right to conduct activities on the internet without being denied the service, unless ordered by court.As time progresses, the battle between privacy-centred techniques and state-sponsored surveillance will intensify. However with surveillance techniques on the rise, the question is whether we want a free society, or a life under total surveillance and control built by the credit-card-like system where you are expected to reveal your identity in order to engage into any activity. The only way we can have any privacy at all is by establishing citizen control over the technology they use.	Thanks to Edward Snowden's disclosures, we know that the current level ofgeneral surveillance in society is incompatible and at odds with humanrights and democratic practice. This situation leads to behaviouraluniformity. The ethical problem stems from the fact that the European Directive is instrumental in continuing to shape a kind of society where citizens have no rights in the digital world.Computer security specialist and cryptographer Bruce Schneier, rightlycommented that "Too many wrongly characterize the debate as'security versus privacy'. The real choice is liberty versuscontrol.". He has also compared the current model of computing asfeudal, one that consolidates power in the hands of the few. Inessence, he categorised technological power into two realms - Big Companies and Governments: "On the corporate side, power is consolidating, a result of two current trends in computing. First, the rise of cloud computing means that we no longer have control of our data. Our e-mail, photos, calendars, address books, messages, and documents are on servers belonging to Google, Apple, Microsoft, Facebook, and so on. And second, we are increasingly accessing our data using devices that we have much less control over: iPhones, iPads, Android phones, Kindles, ChromeBooks, and so on. Unlike traditional operating systems, those devices are controlled much more tightly by the vendors, who limit what software can run, what they can do, how they're updated, and so on."If we want to defang surveillance programs, we need to stop using centralized systems and come together to build an Internet that's decentralized, trustworthy, and free "as in freedom."And the only software that respects our freedom is Free Software. Specifically, free software means users have the four essential freedoms: (0) to run the program, (1) to study and change the program in source code form, (2) to redistribute exact copies, and (3) to distribute modified versions.The idea that we want software to be powerful and reliable comes from the supposition that the software is designed to serve its users. If it is powerful and reliable, that means it serves them better.But software can be said to serve its users only if it respects their freedom. What if the software is designed to put chains on its users? Then powerfulness means the chains are more constricting, and reliability that they are harder to remove. Malicious features, such as spying on the users, restricting the users, back doors, and imposed upgrades are common in proprietary software, and some open source supporters want to implement them in open source programs.A dangerous situation is exactly what we have. Most people involved with software, especially its distributors, say little about freedom—usually because they seek to be "more acceptable to business." The state needs to insist on free software in its own computing for the sake of its computational sovereignty (the state's control over its own computing). All users deserve control over their computing, but the state has a responsibility to the people to maintain control over the computing it does on their behalf. Most government activities now depend on computing, and its control over those activities depends on its control over that computing. Losing this control in an agency whose mission is critical

			on astrategy ensuring citizens to claim and exercise their inalienablerights for freedom of action, both at individual level and at communitylevel. This necessarily implies that EU Citizens can effectivelyinfluence decisions within political, institutional, economic, andsocial systems.Specifically, the top-brass within the executive part of the EuropeanUnion (i.e., the European Commission) must strive in directions thateffectively enables individuals and organisations to affirm theirrights fearlessly against powerful interests through activism,whistleblowing and political journalism; while the technologycommunity focuses on developing effective frameworks to loosen thegrip of governments and tech companies on citizen's computing,including their indiscriminate mass surveillance programs.	https://www.timesofmalta.com/articles/view/20180318/local/government-wants-all-sim-cards-registered-to-help-track-down-criminals.673650 ; 2) https://lovinmalta.com/news/prime-minister-confirms-plan-to-install-facial-recognition-cctv-across-malta). For instance, former special forces operative Michael Yon verified that smartphones are tracking devices that provideactionable intelligence even if location services and GPS aware apps are turned off, or the cell phone itself is shut off.With regards to Lethal Autonomous Weapon Systems (LAWS), Nation States have never been responsible and accountable for casualties, particularly civilian ones. The situation was revealed from information obtained from the Iraq war logs, that revealed 15,000 previously unlisted civilian deaths. (see https://www.theguardian.com/world/2010/oct/22/true-civilian-body-count-iraq). This means that Nation States regularly conceal honest records over the number of wounded civilian, and civilian death counts. These secret programs are operated without any public oversight and outside the limits of constitutions.	society. Left unchecked, this growing copyright-mania will be hugely destructive, particularly for those who choose to use centralized proprietary digital platforms. The only way left for European Citizens is to leave behind the exploitative business modelled school of thought, and unleash the capabilities and potentials of networked systems that are decentralized, uncensorable,privacy-preserving platforms.In relation to education and awareness to foster an ethical mind-set, there occur various problems. For instance the University of Malta IP Policy is being kept secret from the general public (see https://www.um.edu.mt/knowledgetransfer/academicstaff/ippolicy). The policy states that although the copyright shall remain with the originator/s, all computer programmes shall be deemed to be transferred to the University. Specifically, the IP Policy states that the university endeavours to commercially exploit any IP which it owns in collaboration with the Originator. The problem stems from the fact that the term "Intellectual Property" is a made up word meant to disguise the real intention – Ownership of Information andKnowledge, particularly in areas of technology such as computation.(see https://archive.org/details/EbenMoglen-WhyFreedomOfThoughtRequiresFreeMediaAndWhyFreeMedia).Today we use computers to do a lot of things, and universities havestarted to exploit the hard work of students This is what we have todayin many parts of the world, educational institutions putting on cloakssupporting the dissemination of knowledge, to cover the fact that theydo not prepare students to be good members of their communities.			undermines national security.But the most important policy concerns education, since that shapes the future of the European Union. Educational activities, or at least those of state entities, must teach only free software (thus, they should never lead students to use a nonfree program), and should teach the civic reasons for insisting on free software. To teach a nonfree program is to teach dependence, which is contrary to the mission of educational institutions.
Anonymous	Anonymous	Anonymous	Der vorgelegte Text der HLEG ist in seiner Bedeutung kaum zu überschätzen. Den sichtlichen Anstrengungen und dem Ringen um einen gemeinsamen Text ist daher höchste Achtung und Respekt entgegenzubringen. Basierend auf den bisherigen Erfahrungen und Problemen mit KI wird der vorgelegte Entwurf von Ethik-Leitlinien dem sehr komplexen Thema Ethik und künstliche Intelligenz (KI) angesichts der damit verbundenen großen Herausforderungen (noch) nicht gerecht. Angesichts der Bedeutung des Themas ist der Text eine brauchbare Ausgangsbasis, die um wesentliche Elemente (zB mehr Ausgewogenheit im Hinblick auf Risiken inkl. konkreter Benennung und Gegenmaßnahmen), erweitert werden muss, um dem Ziel „Vertrauen" besser gerecht zu werden.Die über weite Strecken einseitige Betonung der möglichen Vorteile des umfassenden Einsatzes von KI unter gleichzeitiger weitgehender Negierung der damit verbundenen Folgen und Risiken für praktisch die gesamte Gesellschaft erscheint kaum geeignet, um den von vielen Stellen und in Medien geäußerten Befürchtungen zu entkräften bzw. bereits absehbaren Konsequenzen mit konkreten Maßnahmen entgegenzutreten. Die großteils unkritische Begrüßung von KI bzw. die Forderung nach deren praktisch ungehemmter Verbreitung und allumfassenden Einsatz sind nicht geeignet, den Bürgerinnen und Bürgern den notwendigen verantwortungsvollen Umgang	Der Text ist als Leitlinie ungenügend, weil es es nicht um die freiwillige Befolgung von Richtlinien sondern um die verpflichtende Setzung von Grenzen in der KI gehen muss. An einem Beispiel ausgedrückt: Wenn Europa nicht zu einem Überwachungsstaat nach ausländischem Muster verkommen will, dann muss irgendwo ein entsprechendes Stoppschild stehen, das KI nicht zur (sozialen) Überwachung eingesetzt wird. Die entsprechende Formulierung greift zu kurz, weil sie sich nur auf staatliches social scoring bezieht („by government") und die nicht-staatliche Seite bewusst und unverständlich ausklammert. Ein klares Bekenntnis zur gesetzlichen, EU-weiten Regulierung von negativen/unerwünschten Auswirkungen von KI ist unbedingt im Sinne der Bürgerinnen und Bürger der Union notwendig und Voraussetzung für das angestrebte Vertrauen, sowohl in die europäischen Gremien als auch in die KI. Diese Schutzfunktion der Experten und der Politik ist auch zum ureigensten Schutz der Forschung notwendig, weil es angesichts der zahlreichen Herausforderungen - nicht nur aber gerade auch im KI-Bereich - äußerst kontraproduktiv wäre, wenn Forschungsergebnisse zur Beschränkung und zum Schaden der individuellen Persönlichkeit missbraucht werden bzw. nicht ausreichend davor geschützt wird. Forschung sollte in diesem Sinne nicht zur Gefahr für die Bürgerinnen und Bürger werden. Die ausreichende Erfahrung, die hoch			Zusammenfassend ist es sehr verwunderlich, dass angesichts der fortgeschrittenen Debatte – zB in den Ausschüssen des Europäischen Parlaments - ein derart unvollständiger und bei weitem nicht auszureichender Text vorgelegt wird. Ebenso ist unverständlich und für die weitere Diskussion äußerst schädlich, dass für die Einbindung der Bürgerinnen und Bürger nur ein englischer Text vorliegt (mit unzureichenden Kurfassungen in anderen Sprachen) und die ursprüngliche Begutachtungsfrist mit einem Monat so kurz angesetzt war, dass an der Ernsthaftigkeit zu zweifeln ist.Der vorliegende Entwurf ist angesichts der von allen Seiten und auch diesem Papier betonten weitreichenden Auswirkungen von KI auf die gesamte Gesellschaft im Hinblick auf den Schutz der Bürgerinnen und Bürger OHNE die angesprochenen fundamentalen Ergänzungen NICHT ausreichend und nicht geeignet die gravierenden Vorbehalte und Folgen zu entkräften oder ausreichend einzufangen.

			der Experten mit diesem Thema zu vermitteln.Ausdruck dieser wenig kritischen Haltung sind zum Beispiel ist zB die Formulierung in den CONCLUSION (S.29) „The AI HLEG recognises the enormous positive impact that AI already has globally“ – dies ist zB mit social scoring und den europ. Werten nicht in Einklang zu bringen, negiert also eine bereits bekannte Bedrohung der europäischen Werte und unterminiert die Glaubwürdigkeit des ganzen Papiers.	entwickelte Gesellschaften mit Forschung und Entwicklung haben, zeigt völlig eindeutig, dass nicht die Begrenzung von Forschung (&E) aber sehr wohl des Einsatzes deren Ergebnisse möglich ist.Ganz konkret wäre es zum Beispiel sinnvoll und notwendig, den vom europäischen Parlament geforderten „menschenkontrollierten KI-Ansatz“ („human-in-command approach“) hier verpflichtend festzuschreiben, ebenso einen verpflichtenden Verhaltenskodex (in der Formulierung des European Economic and Social Committee), der die bloß appellativen Sonntagsreden in konkrete Handlungsanweisungen umsetzt:„Die HLAG FORDERT einen verbindlichen Verhaltenskodex für die Entwicklung, den Einsatz und die Nutzung von KI, um zu gewährleisten, dass während der gesamten Nutzungsdauer von KI-Systemen Menschenwürde, Integrität, Freiheit, Schutz der Privatsphäre und Datenschutz, kulturelle und Geschlechtervielfalt sowie die grundlegenden Menschenrechte gewahrt werden.“Weiters wäre aus zahlreichen Gründen, aber offensichtlich nicht zuletzt zum Schutz der europäischen Soldatinnen und Soldaten, auf die Entwicklung autonomer Kampfmaschinen noch deutlicher einzugehen und deren weltweiten Bann – vergleichbar dem NPT, jedoch OHNE Ausnahmen für Einzelstaaten – anzustreben. Im Hinblick auf die in einschlägigen Diskussionen und Dokumenten zwar oft zitierte aber meist kaum konkretisierte Ethik, muss diese im Text sehr viel deutlicher eingebaut werden (zB unter Verwendung eines EESC-Text):„The development, application and use of AI systems (both public and commercial) must take place within the limits of the European fundamental norms, values, freedoms and human rights. The COUNCIL therefore CALLS for the development and establishment of a uniform global code of ethics for the development, application and use of AI.” Da das Globalziel wohl nicht (sofort) erreichbar sein wird, sollte als Zwischenschritt dieser „code of ethics“ jedenfalls in der Union etabliert werden.		
Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential	Confidential

Roel	Dobbe	AI Now Institute, New York University	SOCIOTECHNICAL RATHER THAN TECHNICAL ROBUSTNESS The two requirements-ethical purpose and technical robustness-do not provide a broad enough foundation to support the "human-centric approach" that the expert group strives for. This can be fixed in an easy and original way: rephrasing the second requirement as SOCIOTECHNICALLY ROBUST. This requirement entails robustness in both the technical systems put in place, as well as in the integration in existing and newly formed social, cultural and organizational mechanisms. A sociotechnical lens recognizes the natural occurrence of more predictable cause-and-effect relationships between technology and humans/society, as well as the complex and harder-to-predict effect that arise due to unknown feedback loops, emerging phenomena and hidden aspects of an AI system (such as energy requirements, hidden labor and maintenance costs). The concept of sociotechnical systems goes back to the 60s and 70s, and recognizes that optimization of each aspect alone (socio or technical) tends to increase not only the quantity of unpredictable, "un-designed" relationships, but also those relationships that are injurious to the system's performance and its human subjects. Embracing sociotechnical robustness as the requirement will allow Europe's leading players in research and innovation to focus on the right problems and increase their lead in developing AI technologies that: (1) are better situated in existing processes and organizations, recognizing the value, knowledge and wisdom of people, and (2) optimize for economic benefit, while proactively measuring and mitigating new forms of harm. As such, the definition TRANSCENDS AND INCLUDES technical robustness, remaining equally relevant in scenarios where AI performs technical and easier to isolate tasks. Note that the HLEG definition of AI (https://ec.europa.eu/futurium/en/ai-alliance-stakeholders-consultation/ai-hleg-definition-ai) is also narrowly focused on the technical fields that contribute to AI. There are a multitude of other disciplines that have long studied AI from a social or sociotechnical perspective that should be seen as integrally part of the AI field, including human-computer interaction, cognitive science, psychology and biology (as pointed out in the definition). Adopting sociotechnical robustness as, will recognize these disciplines and motivate a stronger more impactful cross-disciplinary research and innovation agenda. DEMOCRATIC LEGITIMACY RATHER THAN ETHICAL PURPOSE The other requirement, which says that "development, deployment and use should respect fundamental rights and applicable regulation, as well as core principles and values, ensuring an “ethical purpose”". Having heard from various people participating in the HLEG discussions, the discussion about "which ethics" should be adopted has been difficult and frustrating. Actual ethical principles are highly context dependent and can draw from various schools of ethical thought (e.g. deontological, virtue or consequentialist). As	Following the suggested high-level requirement of democratic legitimacy and sociotechnical robustness (see Introduction comments) it becomes easier to structure Chapter 1, which is now confusing in terms of how it tries to relate "rights", "principles" and "values". Figure 2 is especially unclear, since it is not necessarily the case that principles lead to values, values to rights and rights to principles. This suggested causal cycle will only spur lots of confusion, questions and disagreement. Alternatively, these three concepts can be explained and related in the following way. It makes sense to ground the guidelines in the EU Treaties and Charter of Fundamental Rights. Ensuring democratic legitimacy is only possible if fundamental rights and liberties are respected. A procedural approach centered around impact assessments will then use actionable principles, such as those outlined in the draft guidelines, to consider how existing laws apply that have captured and protected important values that have "emerged" or changed over time, such as privacy or cybersecurity. A value here is defined as something "importance in the life of an individual group or organization", as suggested in the practice of value-sensitive design (VSD) (B. Friedman, P. H. Kahn Jr, A. Borning, and A. Huldtgren, "Value sensitive design and information systems," in Early engagement and new technologies: Opening up the laboratory, Springer, 2013, pp. 55–95). With AI technology quickly evolving and being integrated in many new domains, new concerns will arise around the violation of rights or the conflict of important human values that have not been captured in existing laws. As scholarship in VSD shows, being able to account upfront for all values is impossible. Indeed, scholars from this field note that "values [emerge] whether or not we look for them" (J. Halloran, E. Hornecker, M. Stringer, E. Harris, and G. Fitzpatrick, "The value of values: Resourcing co-design of ubiquitous computing," CoDesign, vol. 5, no. 4, pp. 245–273, 2009). By striving to ensure the second requirement of sociotechnical robustness, an ongoing impact assessment approach will be sensitive to values and human rights violations that emerge in the context in which an AI system is integrated, to track and account for these in the way the system is designed or adjusted. It will proactively engage both the technical and non-technical methods outlined in the draft guidelines depending on the context. As such, values are something to account for, to keep track of and be sensitive to. They will be in conflict across various groups and organizations and require compromise. They will also change over time. Rights are agreements we have fought long and hard for, and which have been historically motivated by our values and the minimal level of dignity. They are a reference frame or standard to measure ourselves and the new systems we build against. The principles then are the mantras we use to proactively assess how a system may or should perform, how it could create new forms of harm and to provide direction to mitigate conflicting values and arising issues in the complex integration of AI systems.			The draft ethics guidelines form a solid start to convene around a set of issues with a very broad set of stakeholders and their views and concerns. As of now, the document is well-grounded in fundamental rights. However, it does not succeed in outlining a clear view on how to turn notions of rights, principles and values into an actionable framework that may be used as a blueprint across various sectors and policy domains. I propose rephrasing the two high-level requirements to "democratic legitimacy" and "sociotechnical robustness" in order to streamline the Introduction and make it more broadly respectful of the crucial disciplines and areas of expertise that the EU has to offer to its own markets and citizens. In addition, I propose to adopt the notion of impact assessments in order to raise clarity around the relationship between "rights", "principles" and "values". An ongoing impact assessment and empowering sector-specific authorities will help develop actionable and scalable approaches. It will also give Europe a large advantage in developing expertise to integrate AI systems responsibly which can be leveraged to grow trade relationships around the world.

such, trying to find consensus around "ethics" itself may prove difficult and unproductive.
Since the ethics guidelines are supposed to work across all sectors and policy domains, it makes more sense to anchor it in the requirement needed for the European Commission to execute its main purpose - being "proposing legislation, implementing decisions, upholding the EU treaties and managing the day-to-day business of the EU" - which is: democratic legitimacy.
The Commission should focus on promoting the internal single market. This includes making sure that all its players (both public and private) behave in a way in which promote business and opportunities, while: (1) respecting existing legal structures, and (2) ensuring fundamental rights and liberties are protected. Democratic legitimacy would motivate empowering sector-specific authorities to oversee, audit, and monitor technologies by domain to meaningfully measure and achieve these two goals.
Each public agency responsible for critical services and infrastructure should be able to organize meaningful public accountability. This implies that trade secrecy and other legal claims cannot form barriers and must be waived, at least to some external auditors in order to adequately assess potential harms and negative consequences.
Corporate secrecy laws are a barrier to due process, which should be guaranteed in any deployment funded by the public sector.
(for more info, see AI Now's Ten Recommendations and 2018 report: https://medium.com/@AINowInstitute/after-a-year-of-tech-scandals-our-10-recommendations-for-ai-95b3b2c5e5)

Using democratic legitimacy also motivates a general solution framework for translating principles into effective ethical practices. Instead of demanding certain principles to be "protected", a more productive means of mitigating negative consequences and maximizing benefits is to ensure that a certain process is followed in which the IMPACT of the technology is ASSESSED on an ongoing basis, involving and responding to all relevant voices. Some key components of an AI or algorithmic impact assessment include: self-assessment of existing and proposed automated decision systems, meaningful external researcher review processes, notices to and consultation with the public before deployment, and provide due process mechanisms for affected individuals or communities to challenge inadequate assessments and harm arising from unfair, biased, or otherwise inadequate system outputs.
(for more information see AI Now's report on Algorithmic Impact Assessments: https://ainowinstitute.org/aiareport2018.pdf)

Seda Gürses			### Remove or heavily amend unsubstantiated claimsWe are concerned that **many of the statements made in this report are not evidence-based or substantiated, and exhibit less of the rigour needed for critical policy-making than the promises and public relations material of industry seeking investment and customers**. These include in particular:- "[AI] presents a great opportunity to increase prosperity and growth, which Europe must strive to achieve"- "AI's benefits outweigh its risks"- "Artificial Intelligence helps improving our quality of life through personalised medicine or more efficient delivery of healthcare services."- "the realisation of AI's vast economic and social benefits" (suggestion: replace helps with might help)- "AI can be a tool to bring more good into the world and/or to help with the world's greatest challenges." (suggestion: replace can with could or might)The report should be capable of making points around the responsible use of technology without reading like a manifesto. Indeed, **such statements are very likely to result in the report losing or lacking legitimacy in the eyes of important stakeholders such as civil society, academic or community groups, as it will lead to confusion as to what the goals, intentions and self-interests of the experts are in promoting AI in such a partial manner**. Of course, proponents of AI hope that "AI's benefits outweigh its risks", but it all depends on where and how it is deployed. We would suggest more conditional phrasing: something like 'We hope that AI will do more good than harm. But in some contexts, the risks of AI may outweigh its benefits.' Any framework for trustworthy AI needs to acknowledge the possibility that in some areas, the use of AI---however trustworthy---would be inappropriate and harmful. This issue is raised later on in section 5, which states: 'due care should be given to what should not be done with AI'. This justified scepticism should be reflected in the statements made earlier on. We touch upon this further below in the section of our comments on Purpose.	### Add missing rights and amend those described.The section on the "Fundamental Rights of Human Beings" **omits the fundamental right to data protection** (Article 8, EU Charter) as something that is "particularly apt to cover the AI field". The omission of this right is shocking: it is conspicuous by its absence and calls into question the motivations of those who did not consider it to be important. It can be noted that data protection law is growing around the world, in countries such as the USA (with the California Consumer Privacy Act) and beyond---it is far from a European blip.**The glossary defines bias but not discrimination**. The definition of bias reduces matters of discrimination to individual bias. This means that discrimination, which is much more complex, is not included as a concern even though that is the focus of most studies currently conducted in the area---which in turn have a limited view on a complex issue that has been deeply studied and the subject of activism for decades. Further, bias (as a form of individual prejudice) is potentially equated with bias in machine learning which is a statistical phenomenon, which might arise from administrative workflows of data gathering, counting etc. Are there experts in the room who have either worked on non-discrimination law, or civil society members who are working on discrimination and equality? If so, which domains do these experts represent (particular beyond the recent domain of 'AI ethics') and have they agreed to this definition of bias? As it stands, we are not convinced there were, which is a heavy subject of concern.There are issues with many of the groups discussed in relation to protection in this report. Considering "children, minorities, disabled persons, elderly persons, or immigrants". **Poverty or class seems to be excluded from this report, as well as attention to historical forms of maginalization**. Given that price discrimination is one of the great promises of AI, the exclusion of class is especially problematic. It is also interesting that for all the talk about discrimination, there is little said about race, religion, ethnicity, possibly all lumped into "minorities". Are there experts in the group who focus on non-discrimination, welfare services, and poverty, and have experience outside of the 'AI' context?	### Rectify the omission of power, purpose, certain peoples and infrastructure**One thing that especially concerns us is an omission of questions of power, the accumulation of power and the control of infrastructure from questions of ethics**. These are among the key concerns in the field as we understand them, and to not include them would risk these guidelines lacking legitimacy, support and uptake, particularly from academia and civil society.#### Omitted peoplesWhen considering whether AI is 'human-centric', it should be explicitly clarified that this also includes challenges that relate in systems key to AI systems' creation, maintenance and deployment to the- the production of hardware and the extraction of resources for this purpose (given the huge importance of powerful hardware to AI systems);- gig workers and deregulation of labour rights through platforms;- the creation of training data, and labelled training data, from overseas workers, such as content moderators, and their labour conditions within the broader value chain.Insofar as AI systems shape sociotechnical systems, business models and incentives, we believe issues are very much within scope of an ethics framework.**In general this AI ethics framework fails to deal with the way that much of AI depends on manual human labor that is situated outside of the EU borders, potentially in countries that have less protection of labor rights**? For example, for content moderation, Facebook has been known to employ cheap labor from Philippines and Morocco, and has been accused of providing limited employee benefits or psychological support, particularly when they are moderating highly distressing datasets. Companies such as Alphabet (through subsidiaries such as Sidewalk Labs) have also been accused of treating smart city incubators, such as Toronto's Waterfront an experimental location that can be used to gather training data from experimenting on citizens, to export to products over the world. Even if these particular companies have put in safeguards in relation to these particular projects, concerns around population wide experiments remain and are likely to be exacerbated, in particularly concerning smaller actors.Is the violation of human dignity in the production and deployment of AI technologies also in the scope of this document? For the production of specialised chips and mobile devices, including the extraction of minerals, often labor conditions that violate human rights are applied. This seems to cause two issues: how far is the jurisdictional reach and the normative scope of these ethical guidelines? The only visible mention of this vast problem is relegated to footnote 12. It would be advisable to promote this point to much more than a footnote.Relatedly, there is a lot of problematic use of the term **European citizen** in this document. Given these global dimensions discussed, and the 'human centric' ambitions of this report, this term should be replaced with a much more generic term that captures both the many individuals, citizens and non-citizens, inhabiting or visiting the EU, as well as those within relevant value chains and affected by these systems who reside across the world.#### Omitted power and centralisationThe issues of the	### Issues with Proposed SolutionsPage 8 suggests that 'tensions may arise between the principles when considered from the point of view of an individual compared with the point of view of society, and vice versa'. **This implies that the only kind of tension is between individuals and society; but perhaps even more important tensions arise between organisations deploying AI and both individuals and society**. In many cases, the interests of both individuals and society will be aligned against those of organisations (whether private or public). Furthermore, many tensions will be between different sections of society or between different powerful interests. Given these different kinds of tensions, the suggestion that there should be an internal and external ethical expert advised to accompany the design may need to be revised or re-interpreted to deal with such cases. How can such an expert truly challenge the business models or practices of the company that pays them?Finally, the legitimacy of any expert's guidance may depend on whether they are truly representative of the interests of affected stakeholders. For those communities who are liable to be harmed by AI, the appointment of an expert who does not understand their needs or represent them in a democratically meaningful way will rightly be seen as a rubber stamping exercise. **The report should emphasise that the expertise needed on social issues and challenges across sectors will often *not* be adequately met or led by data or AI ethics experts, but by those who work primarily on issues in affected communities and are given access to technical expertise.**	
Michael	Veale [and Seda Gürses]	Seda Gürses (KU Leuven) and Michael Veale (UCL)					Comments on the EU HLEG on AI Draft Ethical Guidelines for Trustworthy AI ----------Seda Gürses (KU Leuven) and Michael Veale (University College London)We write as academics working at the intersection of social issues, emerging technologies, and public policy. Seda Gürses is a Postdoctoral Fellow at KU Leuven and incoming Assistant Professor at TU Delft. She sits on the Council of Europe's Committee of Experts onHuman Rights Dimensions of Automated Data Processing and Different Forms of Artificial intelligence (MSI-AUT). Michael Veale is a doctoral candidate at University College London and an Hon. Research Fellow at Birmingham Law School. He is academic advisor to the European Commission's AI Strategy's 'Algorithm Awareness' project.

concentration of power and money in the hands of a few companies that most AI applications currently depend on is not discussed. These include cloud providers, chip manufacturers and platforms. If AI does become as important as the authors of this report hope, what type of democratic control and oversight should be maintained? The tradition of public control of critical infrastructure seems to be largely undermined in this situation. A few of the relevant areas the ethics guidelines omit that they should consider and integrate:- how can small actors easily marginalised by powerful entities engage with or challenge the logics of the infrastructures and decisions from eg large platforms?- which purposes or uses of AI should the public sector retain democratic control over the design and deployment of?In relation to centralised power, there is a large literature and research into the promise of **decentralised systems** which is ignored in this work. Decentralised machine learning and AI (which could be undertaken in many different ways, some more centralised than others) could, in theory, distribute power and decisions over the logics of optimisation of these systems. Who is considered, and who is not considered? Whose interests are at the heart of optimisation systems? Yet from a position of centralisation, decentralisation is unlikely to happen by itself. **The ethics guidelines should indicate that some purposes are too important to centralise and remove decisions about their infrastructural design from the public**, and suggest in broad terms how such purposes might be identified and control over their definition and deployment returned to public hands.Further, in order to develop AI applications, it is often necessary to link disparate databases. When this is done by government entities, this can easily lead to authoritarian results. Furthermore, people may refrain from using one set of government services in fear that it will be linked with information from other services (e.g., classic case being people not using health services when they have problems with tax authorities). What is there to say about the "centralisation of databases" that is typical for AI applications and all rights? We do not believe this issue is out of scope, given the prerequisites of AI systems, it is a closely linked topic.#### Omitted purposes and infrastructuresPhrases such as "**[AI] should be technically robust and reliable since, even with good intentions, a lack of technological mastery can cause unintentional harm**" suggests that harm can only, or will primarily, result unintentionally. This assumes all players are well intentioned. We know from reports about existing tech companies that not all service providers are trustworthy or ethical, nor are their interests always aligned with ethics or positive outcomes for affected populations. Assuming that all service providers are trustworthy could suggest that the only problem is that people are weary of AI because they are scared of innovation, psychologising concerns about negative outcomes due to the application of AI as fear of innovation. This deep assumption that AI is a positive sum game is not substantiated within the document. Indeed, ethical issues are at their most important when incentives are not aligned.**This report needs to acknowledge the limits of trust, and the

fallacy of many interventions that purport to promote it**. We know from privacy discussions that even when there is lack of trust, citizens/consumers/non-citizens may have to keep on using these tools as they become infrastructure. Facebook has lost much trust, but has not shed many users or advertisers. The authors may learn much from the history of privacy and trust in thinking about AI systems. It may be a welcome addition to draw on existing scholarship and experiences. A similar lesson can be learned from the history of 'notice and consent', which has not provided effective control to individuals, who have become resigned to using systems they cannot challenge. A similar situation is likely if individual remedies, such as transparency or 'explanation' are relied on in machine learning, rather than empowering groups, regulators, civil society and/or journalists. **Encouraging 'Trustworthy AI' is neither here nor there if individuals are forced to participate in systems whether they are trustworthy or not**.**The report also focuses on developers and implementers, which is shortsighted, and assumes such parties have full control over these systems' development and deployment.** Unfairness may occur after deployment, due to a third party use of AI based services, or due to interaction of multiple services in an environment. The assumptions that developers can foresee these issues, will have the means to address, and can effectively address all these outcomes is a narrow view on how these systems are or should be governed. Governing systems up-stream is welcome, but consideration should be given to how models are increasingly being traded and moving around the world, integrated in products and value chains and sold onward.Currently, companies that develop AI are also very much bound to platform models, be it by virtue of using cloud services, or using platforms like Facebook, Apple, Amazon or Google. This means the production and deployment of AI is subject to a complex governance structure where being ethical for developers may simply depend on how ethical the platforms are. What does this document say about the different division of responsibilities and delegation of responsibility and complex ways in which structural disadvantage may arise from these systems in indirect ways? Again, it would be very helpful to have people who are articulate in matters of structural discrimination, as well as (political) economy in the room to be able to address these issues appropriately.Overall, the document lumps all AI together, and does not consider the role of risks and potential negative outcomes in evaluating whether AI should not be applied in some contexts. Even in the US, where regulation lacks behind, there is by now common understanding that some systems shall not be deployed (e.g., facial recognition in policing). Some systems are too dangerous, are too risky especially for vulnerable populations. This is independent of whether the algorithms are fair. Giving information to these populations and making AI systems auditable is not sufficient when such concerns are present. If there is an ambition to build ethical guidelines, then these should also provide guidance to stakeholders as to when it is not appropriate to move decisions or the management of environments (both

natural and constructed) to AI based systems. Phrases such as "[t]he aim is to foster a climate most favourable to AI's beneficial innovation and uptake" bring a strong inevitability assumption here with respect to AI being the technology and having to be applied. It seems a very narrow view on ethics, especially if the experts shy away from providing normative guidance on when there is a red line in applying AI. It is also odd that the main reason for applying ethics is to improve innovation and uptake. This suggests that ethics is a tool to enable innovation and uptake, leaving readers concerned about whether ethics will kick in when innovation and uptake has negative consequences.**Parts of the report appear to discuss concerns that are present in business models today as if they are future concerns**. For example, the guidelines state "people with the power to control AI are potentially able to manipulate humans on an unprecedented scale". One could argue that most AB testing is a (for the industry acceptable) form of manipulation. So, where do the experts draw the line? Similarly, the section on 'Normative and Mass Citizen Scoring' should be mindful that this extends far beyond public authorities to the scoring of individuals and employees by private actors, such as workers on platforms.Discussion of **covert AI systems** should be given a broader, systemic scope. We do not believe that it is only when an individual is interacting with an AI system that covert systems are concerning. If AI is shaping one's environment, e.g., when Waze routes cars through neighborhoods, we believe this has grounds to be described as a covert AI system, particularly insofar as it falls outside of meaningful democratic accountability mechanisms. A similar situation can be said for when AI based services are running experiments to get better outcomes (e.g., for increasing profits)? It is insufficient to only refer to "interacting with AI" in a way that does not cover all of these cases.When Waze navigator reroutes cars out of traffic to surface roads, it is exemplary of an environmental effect that benefits the users while externalizing risks and costs to those living on those surface roads, as well as the municipalities managing those roads. In most cases, the residents of those roads will not be able to intuit exactly what is going on and why these cars are overloading their streets. This happens because service providers make heavy use of public infrastructure without further accountability or responsibility. Companies like waze have been known to not respond to requests from residents, but also municipalities or road services. In a sense, they can't since this would constrain the scalability of AI based services. Would such cases fit in the principle of explicability and if so how?### Human Centrism is not EnoughThe report argues that the focus of AI should be increasing "citizen's well being" [sic]: however, AI is likely to be applied just as much to non-citizens, to environments, to animals etc. This humanist approach with a focus on citizens is too limited and leaves all others beyond ethical boundaries. We can see how focusing on human rights may cause this bias, but considering that AI is just as much about transforming environments, e.g., smart cities, health etc., focusing only on human well being without taking their

						relation to these environments into account is likely to unduly limit ethical concerns to AI applications that directly interact with people. For example, how would this ethics document apply to precision agriculture, especially if that technology produces more produce, at lower cost, but at the cost of environmental destruction. These issues certainly have ethical dimensions. This is touched upon later on page 9 regarding environmental effects, but it unclear how this links to a vision of 'human centrism'. Furthermore, Footnote 11 is problematic, and appears to show a shocking ignorance for the long history of research into sustainability and its definitions (see e.g., the discussions of Weak Sustainability vs Strong Sustainability; the notion of sustainable development, among many other areas.). We suggest this part be reviewed by an expert from the environmental policy field as not to lose legitimacy within these important domains.		
Pablo	Ramirez	eurokompras	The starting point of the guidelines should focus on the social good.	The data that will be used to develop the AI must be obtained in respect of the privacy of the citizens			we have to develop and establish concurring criteria that define what you can do and what you can not do with the AI in addition you could establish an institution that evaluates and controls the fulfillment, even the AI itself could evaluate itself	The AI can provide new ways of approaching problems and meaningfully improve people’s lives. With AI, we have another tool to explore and address hard, unanswered questions. What if people could predict natural disasters before they happen? Better protect endangered species? Or track disease as it spreads, to eliminate it sooner? AI can help, but it’s not a silver bullet: tackling these questions requires a concerted, collaborative effort across all sectors of society. We must define the priorities to develop solutions for the common good, it will be difficult but together we can do it

Anonymous	Anonymous	Anonymous	Comments have been captured within the following section for this.	(This also includes comments for the above introduction section) One of the highest level comments that we make is the following:The AI Ethics guidelines should be an evolving/iterative and participatory process. It should include multiple stakeholders and have a system of regular review. Licensing might be something that helps to enact the legal code and bring about accountability and attribution, for example digital signatures for versioning AI systems and then identifying individuals who are held accountable, similar to a Data Protection Officer that is nominated within an organization and who is responsible under the GDPR. Also, given that laws differ across many nations and regions, that needs to be something that the guidelines have to provide some guidance on. For example, in the region of Quebec, an engineering degree comes with a code of ethics, that is perhaps something to be considered when it comes to software engineers working on AI that adhere to a code of ethics or perhaps are bound by some certification. Another thing that the guidelines should acknowledge and address are the costs of compliance from a legal perspective, for example, as in the case of GDPR, where it might place significant burden on organizations that have limited resources. Also, there might be some use in distinguishing the implications in different categories of use e.g. civil vs. military. An approach that we found would be useful in the development of these guidelines and subsequently their implementation is that there be a quasi real-time collaboration - regulators and policymakers work hand in hand with tech industry that can work together to achieve the following:-team effort - merger of effort - further than collaboration - -reward for stringent rules by licensing strict tech players -partnership on AI -trust in AI is critical to industry players and thus requisite standard of protection and adequate level of care is beneficial collectively - standards of organization of AI-Tools (based on use cases or complaints or precedents) - -industry commitment too complicated due to all fields all industries application of AI-2-3 sets of standards rather than the plethora of non-binding declarations, etc. that are out there at the moment. Some other considerations to think about:-legal mechanisms may be insufficient - too fast paced for human, invisible to the real world, and regulatory sampling may be insufficient - technical mechanisms may be optimal (eg. privacy by design) - unresolved in terms of coding values or morals within AI -standardized production of ai v. design of ai (i.e. argument around how most AI system design today is artisanal compared to the industrialized approach we see in other fields) Perhaps, checklists that help developers and designers navigate might be helpful which could be in the form of yes/no questions and a decision tree on what to do. From an autonomous action perspective:-What is the adequate framework - large organization taking facets (religion, tech industry) - gathers expertise - council of educators - see AI as a child - global community of educators - -law alone is insufficient - adequate enforcement is critical-group leaders in an industry setting an example calling out the bad players (market or industry self-regulation) -gap between the enforcement -compliance by design? -You can have an ethical product,	As a meta-level suggestion, we begin by acknowledging that the 10 requirements for Trustworthy AI are currently listed in no particular order. We understand that they're all important and the difficulty in prioritizing them. However, a side effect of that is that none of them seem particularly important. This is likely unrepresentative of the reality of ethics in AI, just as it is in the ethics of everyday life. To that end, we'd suggest establishing a proper hierarchy such that teams with limited resources know which aspects to focus the bulk of their energy and time on. Pareto's Principle is a useful mental model here: 80% of the results will likely come from 20% of the inputs. Given that there will be domino effects, we'd recommend pointing out for example which 3-4 lead elements or requirements, if implemented well, will make all the other ones fall into place as a side-effect, or at least make them easier to achieve.11 more suggestions from a bird's eye view of the section on realizing trustworthy AI:Introduce the idea of a participatory system where the AI isn't simply adhering to a black box, but is actually communicating with its user about how best to help them. It may not be enough to simply educate those building and implementing AI - educating the general public about how best to interact with it may play a crucial role in the responsible development of trustworthy AI - as with any business, it's important to build customer-centric products and services.Some more details on the standardization process and how such certifications could be implemented and who'd hand them out would be useful, as well as how to keep these organizations' incentives aligned with the people's. Think about how we can create a culture where it is not stigmatized for people to opt out of an AI-enabled world without inviting scorn or being called a luddite. When AI systems are built transparently and people fully understand what the implications of participating in that are, AI wouldn't truly be supporting people's autonomy if they don't actually have the option of saying no due to social pressures.Some ideas or predictions about the barriers that different government institutions may face in attempting to implement these ideas may serve as a useful pre-mortem.What methods of enforceability can citizens and the government use to hold those implementing AI systems, to these requirements and ideas?This section is called "Realizing Trustworthy AI" but it is not very technical or practical, i.e. does not bear relevant relation to the people or ways in which AI will be "realized". There should be specific guidelines, perhaps in the format of checklists or flowcharts, for developers of AI systems. Consider combining "Data Governance" and "Respect for Privacy", and creating 3 new categories about "Technology and Development", which should address technical best practices, and how they should be established in joint efforts between policy makers, researchers, and other stakeholders. These practices should be subject to revision. The 2nd category we recommend is "'Education", of the general populace and specific stakeholders about statistics, machine learning, and responsible use, of researchers about relevant ethical frameworks, policy, and best-practices, and of people using specific technologies about how to use them properly and safely. The	For this section, roughly working along the lines of 1) What's missing from the section? 2) How it can be made more accessible to the intended target audience? Here are the insights: 1) What are the missing elements in this section keeping in mind the stated purpose of the section?Contextual approaches make a 'one-rule-fits-all' assessment more complex. For the insights to be practical and applicable, context is crucial. For example, for governing the autonomy of AI, approaches to be used in self-driving vehicles vs. diagnostic systems in medicine will differ by a lot. An overall comment on the guidelines provided in the document is that they are very high-level and not very actionable. AccountabilityLacks examples of what accountability represents and how that may be modeled in real world cases. What is accountability in the context of AI? (providing a definition here would be very useful). There is always a notion of failure on the part of the system and learning outcomes from those is a trial-and-error process which is ultimately valuable in improving the systems.When asking 'who' is accountable, what are we referring to - i.e. human-level entities, system-level entities, etc.There are some other aspects that come into play when talking about accountability like compliance, service agreements, perhaps even letters of intent that define boundaries and scopes for the accountabilityTwo types of accountability: one towards the client, external, one internal within the companies' boundaries towards the people working on building these systems.What is also important is to define critical stages of 'wrong' - and having procedures accordingly to address them. We should also be able to in accordance measure compliance at various levels. The section could do better by having more specificity regarding processes/mechanisms that help to achieve the accountability objectives.The measures of accountability should be tied to the impact of AI - for example, the use of tools like algorithmic impact assessments, sample ones have been produced by the Treasury Board Secretariat of Canada and the AI Now Institute.Data governanceWe need to be able to maintain the integrity of the data and simultaneously address questions on provenance of data: Where does the data come from, who owns it? In a governance context there is a need to have an authority to decide on these issues - Who is this authority? Who has the authority? Who should have the authority?There needs to be a degree of transparency and user-level awareness of intent of use regarding data. Specifically that intent should be explicit such that users know what they are consenting to. Right to forget, right to data suppression. How do we have both informed consent and be able to remove our data from a system without breaking the system if it has already used our data - specifically in a machine learning context where the learned representations have used that person's data in the training phase. A possible solution is re-training the model but that is too costly when large datasets with expensive training cycles (in terms of computation budgets) are involved.Design for allThe concept as expressed in its current form is difficult to understand and needs to be articulated better - it also stands in conflict with other values as outlined in another part of this	I believe that it is worthwhile to highlight the process through which these insights have been generated and the AI Ethics community (https://montrealethics.ai/community/community.html) in Montreal that has played a crucial role in the success of the preparation of these remarks. There were two events (https://www.eventbrite.ca/e/ai-ethics-european-commission-guidelines-feedback-session-tickets-54494772331 and https://www.eventbrite.ca/e/ai-ethics-final-session-european-commission-guidelines-feedback-tickets-55244970193) held over the course of a month that captured insights from a diverse set of people comprising the AI ethics community in Montreal that I founded. The process that is followed in capturing and eliciting the insights is highlighted here at the beginning of the article: https://medium.com/montreal-ai-ethics-institute/ai-ethics-inclusivity-in-smart-cities-6b8faebf7ce3 The community strongly believes that some of the most meaningful and impactful solutions are going to arise out of an inclusive and participatory process that involves people coming from diverse backgrounds and not just experts that are sourced from traditional avenues. It also serves another purpose in educating and empowering even more voices in being able to contribute to the discussion on building trustworthy AI systems that will ultimately result in systems that benefit everyone in every sense of the word.

but it can still be used unethically - how to ensure good use of AI -hard code acceptable behaviour or use of AI within AI as a first line of defense possible? Large burden on tech players - society must also be responsible - hence for the framework - should people who are knowledgeable and who are developing the AI have a higher responsibility or care. -Second line of defence: regulatory enforcement, supervision and monitoring? But how far-Third line of defence - manual kill switch -AI DNA - preliminary basic code imposed for all algorithms to v. new DNA - possibe? There yet? Want this? Countries may disagree. Behavioural OS (operating system) to define - perhaps not far removed by the realm of the possible - limit the possibilities of the AI - Needs to be iterative process. But no human values embedded - the security safeguards within OS may be necessary. - Core AI behaviour to be standardized.From a professional designation & minimum standard perspective:-determinism - important part of AI-however, many technicians have developed skills in ai programme outside of the traditional education system + peer review system validation-invention and creativity v. law and accountabilitytoo dangerous a tradeoff - over-reliance, overconfidence, carelessnessTaking the present context into consideration - -ethical certification - - regulators may not have the technical expertise or knowledge to regulate this - this may be a problem, however, Declaration of Montreal - multidisciplinary - teamwork - non-siloed collaboration - so lack of expertise of politicians and regulators may be mitigated by consulting experts in the field and the general population at large. - Bridging the Gap - concern that lawmakers get together to create policy without expert understanding or knowledge, stifling innovation in the field - -traditional democratic consultation processes - - corruption of regulators or policymakers or lobbying by bigger participants are other concerns that we need to be thinking about.Other general comments about this section:From a public awareness and education to understand what people wantDefining Good is very important and for whom?Collective Good - But can't forget minority eitherEducating AI to follow human experiences, at least in the positive sense.Global consensus on inappropriate behaviour is not something that we have and hence contextual sensitivity is important.Educational awareness is incredibly important - so strong supportive education system and support system must be put in placeEthical ecosystem - provides a framework where programmers can propose new ideas (eg. Apple and their strong commitment towards privacy)Commitment: A global commitment is key, to avoid race to the bottom and ensure at least minimal protection Right to science ? Everyone has the right to benefit from science.Access to data? How do we achieve informed consent? Should it be per decision? bioethics moved from individual interactions to public, population level ethics Can a population consent? e.g. of doing genetic studies that happen in a region known to have a lot of people that have a certain genetic disease - insurance premiums will be driven up even if you haven't participated in the study but fall under the demographics for which the study

3rd category we recommend is "Mitigation and Restitution" i.e. what to do if things go wrong, discuss responsibility and accountability, and to have built-in systems and fallbacks for what to do. Perhaps these concepts could be merged in to AI Governance.For all sections, but this in particularly, the work would benefit from a case-study approach to make things concrete, especially in the cases where there are potential conflicts between objectives. For example, if technology is supposed to be developed so that anyone can use it anywhere, how do we balance the safety and security of that technology? If I have the right to be forgotten, can I have data deleted that is necessary for an ongoing or future court case? See "potential conflicts" below for these and more examples.The list would benefit from more (and more consistent) structure; some sections have subheadings and these are helpful, but others do not and they could/should. Several of the topics are heavily overlapping and could be combined, or at least structured in a hierarchy or venn diagram, and cross-references/links about how parts relate to each other should be added, as done in the Montreal Declaration on AI ethics.References to other relevant guidelines and documents are almost entirely lacking, as well as references to relevant types of guidelines from e.g. nuclear physics, pharmaceutical testing, etc.Below is a more direct 1:1 feedback list in terms of missing nuance.Accountability: Independence of mechanisms of accountabilityContinual reassessment Interpretability Querability Meeting standards (detailed, application-specific), certificationDataDoesn't address who has access to the data / whether it can be shared and if so with whom and under what circumstances or agreements.Data expiry/right to be forgottenCan't gather data just in case / unnecessarily unless you meet a threshold volume of data where the marginal contribution of an individual is negligible and therefore cannot identify individuals, but also have an opt-in ability to be identified (e.g. I have a rare disease and want you to be able to study it).Integrity/encryption of storageDesign for allConsider a more nuanced look at the regulatory burden andprice of AI, or acknowledge that this could hold up progress to the point of being disregarded.Discuss the outcomes of technology; this section talks only about the technology itself but should include something about the benefits accruing to all / redistribution of wealth created by the use of the technology.Respect for human autonomyMention the ability to meaningfully opt out (as discussed more deeply earlier)Mention mitigation and responsibilityNon-DiscriminationDoesn't mention any of the extensive literature on fairness and discrimination; e.g. positive discrimination is sometimes desirable, and there are different definitions of fairness which lead to very different outcomes and recommendations (e.g. equal outcomes vs. equal opportunity) - tell the full story.Discuss the difference between malicious vs. systematic discrimination.Given that all data can be biased, it doesn't make sense to say that practitioners must "prune away" biases or not use biased data; rather we suggest to say that actions cannot be taken which could conceivably reinforce or encourage existing but ethically misguided

submissionThis concept might also be scoped such that design for emphasized If its made for general, public use but if there are specific uses for building a product there might be trade-offs in making it more effective and useful for the intended user base. Differentiating between general purpose technologies and specific use-cases for sub-segments is a factor in certain cases. How do create adaptable designs that are based on 1) how users live their lives and 2) representative datasets?How do you cater for technologically challenged segments of the general population? The system has to be equitable for those it is targeted for, which is not necessarily 100% of the human population. The data subjects that are more willing to give their data may not be representatives of all users, results in self-selection bias in the AI. We see this happen involuntarily, for example, in using the Reddit 2 billion comments dataset for natural language, we can't build ML systems that are representative because internet users are just a subset of the human population. Reddit users are a subset of the internet users and even within Reddit users there is a small percentage that contributes the largest amount of data vs. the others. By being very general, the guidelines are very difficult to implement concretely. There also need to be some considerations towards how these might be put into practice for resource-constrained organizations - e.g. startups, non-profits, civil society organizations, etc. What happens when the cost of non-discrimination is very high and acts as a burden for the company as a whole eg. elevators in the Paris metro, medical imaging AI that doesn't cover certain minorities The degrees and extents of each of the points should be adjusted to serve different contexts. Certain technologies pose higher levels of threats to privacy, safety.Grading rather than binary approach is probably preferable to be realistically usable. They are not actually binary, but more "idealistic", but not necessarily enforceable. Guidelines can serve as a first step for legislation and for improving existing legislation. They can be used as a consideration; the law already has general principles that governments rely on. These guidelines can do more in terms of being actionable. Governing AI autonomyFor bullet point 1: "each stage" seems vague and needs to be articulated more clearlyFor bullet 2:"Self-learning" should be replaced by more precise technical language which ultimately helps it in being more actionable.Suggest addressing this issue: AI systems can cause effects which continue after the system has been shut-down. For instance, the damage caused by an AI which promotes a fake news story cannot be undone by shutting down that system. Is there a plan for pre-emptively intervening before an AI system causes such problems, identifying the potential for such problems and/or managing them after a system is shut down?Is there a backup system which can safely replace the AI system if it needs to be shut down?Suggested bullet point: Is this system capable of learning "online" (i.e. after it has been deployed)? If so, is there a plan for recognizing significant changes in the system when they occur? Will there be periodic assessments of what changes have occured?Non-discriminationAlready has a legal basis from which relevant sections

was donecan use the core principles in public health ethics Apply public health ethics to expand analysis to macro/ population level.Move from individual point of view to system point of view Fear of negative impact when individuals can be identified. Prevent re-identification. Make triangulation unlawful. Avoid mosaic effect.Trade-off between the strength of the privacy and the usefulness of the data. (differential privacy) Can people make the distinction between data governance and AI algorithm?Education and public competence building are important toolswhen talking about “doing good” - what is it exactly? Jehova witness - saving the soul vs saving the child’s life and refusing blood transfusion religion is completely missing from the report - from the doing good and do no harm - this can vary from one person to another based on their religious beliefs which are important, perhaps even when thinking about what good means to them.What is “do good” ? “do no harm”? - Beneficence and empowering the person. Right of the individual to challenge AI decisions.If impacted by an automatic decision there should be access to a “human path” around it.Where should we use AI anyway? Avoid bias when AI isn’t really needed.Humans don’t even agree on what’s good, what’s “do no harm”!Opt-out is good, but can we go against a decision ? Can we reverse a decision? What about compensation? Autonomy - Can you go against something that is objectively good for you?Regarding medical diagnostics: doctor/patient privacy. With AI, information might be shared across many use cases for training purposes unless there are strict guarantees that are provided. Can AI influence people in making them different decisions than they would have done without AI? If you opt-out, the document doesn’t mention alternatives that people can use in that situation today the flow is that AI systems are there to support humans in making decisions Is there an opposition to AI making systems for humans because of the principle of autonomy? or because we believe that AI systems make decisions that are biased and worse than humans? How do all the stakeholders align? Who gives direction at a macro level over all the organizations? Something that can be done to make the report more effective is to have a focus on case studies, storytelling and infographics. Don’t provide a PDF but a dynamic website where people can choose their paths, videos. Fear needs to be transformed into fun as people learn about this and are encouraged to apply this to their research and work. Another option could be to run workshops that guide people on these topics. Provide cookie-made workshops that companies/groups can choose to run locally. These could be provided by perhaps government and industry leaders The document could be split by public and private industry. It should also be separated out by different domains that might be affected by this. Create various versions of the document to match various domains. Match their vocabulary. Adapt the text and presentation to the audience in order to make it accessible and understandable to all. Offer a forum where people can ask questions and get answers. LAWS need to be looked into a lot more before publishing the final version of this report. There needs to be more added on privacy by design and on the biases. (e.g. predictive policing)Human oversightDiscuss the problem of “relevant” or “real” human oversight, i.e. it’s not real oversight if the human just has to push a button without understanding or adequately considering the effects of that action, but it may be beyond the scope of a human to understand the decisions an AI makes.PrivacyCould be combined with “data”, or if it is different, it should discuss the ways in which it is different. (e.g. private access to the outputs of an algorithm..?)Should mention differential privacy and the substantial literature on when and why (and when and why not) it is an applicable tool.RobustnessSome information should be added about calibrated confidence; that AI systems should understand their own limitations and be able to accurately express and predict their own uncertainty.SafetyShould mention the restriction of certain technologies for safe use, and require proper training and education about how to use advanced and/or potentially dangerous technologies properly.Be conservative about the deployment and access to potentially dangerous technologies, and work to establish accurate understanding of what is and isn’t potentially dangerous.Limiting resources and autonomy of independent systems and ensuring both auditability and meaningful human oversight should be discussed.TransparencyProactive disclosureAuditabilityFinally, it is important to point out that there may be potential conflicts between some of the 10 requirements. These may have to be resolved through experiments tailored to each unique situation. We explore one example more deeply to spark ideas as to other conflicts.Human Oversight vs. Non-DiscriminationWe understand the need for human oversight with regards to autonomous systems: We want people to be in charge of major decisions when ethics are involved - that is clear. However, things become a little more murky when we realize that sometimes having a human in the loop may actually introduce bias into the “clean room” of an intelligent machine’s black box. This is because human beings are prone to a number of systematic biases, as explored in the literature around behavioral economics and social psychology. It is difficult to build non-discriminatory AI systems when the ultimate decision-maker (the human in the loop providing oversight) is herself prone to making decisions that may be deemed to be discriminatory in nature.Safety vs. Design For AllRespect for Privacy vs TransparencyData Governance vs. Accountability Human Oversight vs. Robustness Transparency vs. RobustnessAccountability vs. Respect for Privacy should be cited and refined upon. The question is how do you interpret non-discrimination in the context of AI?The Amazon HR AI example potentially discriminating against certain candidates but it was only done in a test context and highlighted in the mediaWhat happens when the cost of non-discrimination is very high - especially for resource constrained organizations as highlighted in the comments for section 3.Respect for PrivacyAlready has a legal basis which can be tapped into to come up with more concrete suggestions on how this can be put into practice. Respect for Human AutonomyWe suggest a change to the title of this section to: Respect for (& Enhancement of) Human AutonomyMental integrity is a strong expression to qualify simple marketing decisions. Should provide support to the user but not dictate their behaviour.E.g.. when using a self-driving that uses a GPS which tells you that you cannot go right, there should be an ability to “manually override” the algorithm. Even today, do you really “choose” to follow the GPS or are subconsciously giving away autonomy to a system that guides you through a certain path that might not be optimal for you on the individual level but is optimal for the entire traffic grid and that is not made explicit to you. The role of “head of pricing” or “head of data” or “head of compliance” - explainability is key and should not rely on sales people or individual customers to observe. The burden cannot be placed on people that don’t have all the information nor the resources to be able to make these decisions.Suggested bullet point: What methodology was used to assess whether and how the AI system can influence users’ decision-making or beliefs?Suggested bullet point: If the AI system influences users’ decision-making or beliefs, is the influence of a form that users would endorse, e.g. consider informative, rather than manipulative?RobustnessThis sections is a lot more developed and is more understandable by technical people. It is how we think the rest of the document should be written.Suggested bullet point: Has the system been tested by “red teams” (who have the goal of demonstrating flaws or problematic behavior in the system) akin to the work that is done in the domain of cybersecurity.SafetyFor the 3rd bullet point: they should also ask: “Could interactions with other automated decision making systems cause emergent safety concerns?”Suggested bullet point: Would this system impose safety risks on people other than its intended users? Is there a plan for managing such externalities? Suggested bullet point: Has the domain of safe operation of the system been assessed? How? (e.g. are the assessments based on testing (in the real world? in simulation?), theoretical justifications, or some other method?)TransparencySuggested bullet point: Have the limitations of the system been assessed, technically? How?Other general remarks on the section:Many of these guidelines already have a legal basis; these guidelines are simply trying to provide a specific context to interpret legislation in the new context of emerging AI solutions. Though this section pertains to “assessment”, it says very little about how concretely audits are to be achieved. Is enforceability an issue? There is also the

				notion of consent.				question of being able to prove that a company is discriminating against its users, an example being Netflix supposedly showing slices from movies as the thumbnails with African-Americans to users who they thought were supposed to be African-Americans. AI uses for marketing can easily delve into discriminatory practice. Before releasing the algorithm, there should be some compliance steps. In you take a random selection, all people should be equitably represented in the samples. The intended audience for these guidelines is a more general audience which it definitely caters to as it is but it will benefit from having more depth to allow for it to be practically useful to policy makers, technical members, etc. In human life we check ourselves, the algorithm doesn't check itself, therefore this needs to be addressed explicitly. The data fed to the algorithm, not the algorithm itself should not be discriminatory, i.e. the source of the biases is the data rather than the algorithms themselves.Considering proxy discriminatory bases would be key, e.g. attributes like zip codes strongly correlate with ethnic origin in places like the United States.		
Pradyot	Sahu	3innovate							1-An AI ethics guideline like this may be a first step. A GDPR like regulation may be created and implemented. 2-Autonomous AI weapons should be banned like the chemical weapons. There should be international level discussions immediately to ban autonomous AI weapons. 3-Since the field of AI is changing very fast, new things previously unknown will emerge. This guidelines document should be reviewed again and again after few years to incorporate new changes.