European Commission

The European Commission's

HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE

DRAFT ETHICS GUIDELINES FOR TRUSTWORTHY AI

EXECUTIVE SUMMARY

Working Document for stakeholders' consultation

Brussels, 18 December 2018

DRAFT ETHICS GUIDELINES

FOR TRUSTWORTHY AI

High-Level Expert Group on Artificial Intelligence
Draft Ethics Guidelines for Trustworthy AI

European Commission
Directorate-General for Communication

Contact Nathalie Smuha - AI HLEG Coordinator
E-mail CNECT-HLG-AI@ec.europa.eu

European Commission
B-1049 Brussels

Document made public on 18 December 2018.

This working document was produced by the AI HLEG without prejudice to the individual position of its members on specific points, and without prejudice to the final version of the document. This document will still be further developed and a final version thereof will be presented in March 2019 following the stakeholder consultation through the European AI Alliance.

Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of the following information. The contents of this working document are the sole responsibility of the High-Level Expert Group on Artificial Intelligence (AI HLEG). Although staff of the Commission services facilitated the preparation of the Guidelines, the views expressed in this document reflect the opinion of the AI HLEG, and may not in any circumstances be regarded as stating an official position of the European Commission. This is a draft of the first Deliverable of the AI HLEG. A final version thereof will be presented to the Commission in March 2019. A final version of the second Deliverable – the AI Policy and Investment Recommendations – will be presented mid-2019.

More information on the High-Level Expert Group on Artificial Intelligence is available online (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

EXECUTIVE SUMMARY

This working document constitutes a draft of the AI Ethics Guidelines produced by the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG), of which a final version is due in March 2019.

Artificial Intelligence (AI) is one of the most transformative forces of our time, and is bound to alter the fabric of society. It presents a great opportunity to increase prosperity and growth, which Europe must strive to achieve. Over the last decade, major advances were realised due to the availability of vast amounts of digital data, powerful computing architectures, and advances in AI techniques such as machine learning. Major AI-enabled developments in autonomous vehicles, healthcare, home/service robots, education or cybersecurity are improving the quality of our lives every day. Furthermore, 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.

Having the capability to generate tremendous benefits for individuals and society, AI also gives rise to certain risks that should be properly managed. 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**. 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. **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.

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.

These Guidelines therefore set out a **framework for Trustworthy AI:**

- **Chapter I** deals with **ensuring AI's ethical purpose**, by setting out the fundamental rights, principles and values that it should comply with.
- From those principles, **Chapter II** derives **guidance on the realisation** of Trustworthy AI, tackling both ethical purpose and technical robustness. This is done by listing the requirements for Trustworthy AI and offering an overview of technical and non-technical methods that can be used for its implementation.
- **Chapter III** subsequently **operationalises** the requirements by providing a concrete but non-exhaustive assessment list for Trustworthy AI. This list is then adapted to specific use cases.

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).

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.

Importantly, these Guidelines are not intended as a substitute to any form of policymaking or regulation (to be dealt with in the AI HLEG's second deliverable: the Policy & Investment Recommendations, due in

May 2019), nor do they aim to deter the introduction thereof. 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"**.

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.

Finally, beyond Europe, these Guidelines also aim to **foster reflection and discussion** on an ethical framework for AI at **global level**.

EXECUTIVE GUIDANCE

Each Chapter of the Guidelines offers guidance on achieving Trustworthy AI, addressed to all relevant stakeholders developing, deploying or using AI, summarised here below:

Chapter I: Key Guidance for Ensuring Ethical Purpose:

- Ensure that AI is **human-centric**: AI should be developed, deployed and used with an "**ethical purpose"**, grounded in, and 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.* This is crucial to work towards **Trustworthy AI**.
- Rely on fundamental rights, ethical principles and values to prospectively evaluate possible effects of AI on human beings and the common good. 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.
- Acknowledge and be aware of the fact that, while bringing substantive benefits to individuals and society, AI can also have a negative impact. Remain vigilant for areas of critical concern.

Chapter II: Key Guidance for Realising Trustworthy AI:

- Incorporate the **requirements for Trustworthy** AI **from the earliest design phase**: Accountability, Data Governance, Design for all, Governance of AI Autonomy (Human oversight), Non-Discrimination, Respect for Human Autonomy, Respect for Privacy, Robustness, Safety, Transparency.
- Consider technical and non-technical methods to ensure the implementation of those requirements into the AI system. 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.
- 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. Ensuring **Traceability** of the AI system is key in this regard.
- 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. Trustworthy AI can also be included in organisations' deontology charters or codes of conduct.
- 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.

- 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.
- Ensure a specific process for **accountability governance**.
- Foresee **training and education**, and ensure that managers, developers, users and employers are aware of and are trained in Trustworthy AI.
- Be mindful that there might be fundamental tensions between different objectives (transparency can open the door to misuse; identifying and correcting bias might contrast with privacy protections). Communicate and document these trade-offs.
- Foster research and innovation to further the achievement of the requirements for Trustworthy AI.

Chapter III: Key Guidance for Assessing Trustworthy AI

- Adopt an **assessment list** for Trustworthy AI when developing, deploying or using AI, and adapt it to the specific use case in which the system is being used.
- Keep in mind that an assessment list will **never be exhaustive**, and that ensuring Trustworthy AI is not about ticking boxes, but about a continuous process of identifying requirements, evaluating solutions and ensuring improved outcomes throughout the entire lifecycle of the AI system.

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. It strives to facilitate and enable **"Trustworthy AI made in Europe"** which will enhance the well-being of European citizens.