UCLouvain

CRIDES

Centre de recherche interdisciplinaire
Droit, Entreprise et Société

GENERAL REMARKS CONCERNING THE ASSESSMENT LIST

By Anne-Grace Kleczewski[1] and Prof. Alain Strowel[2]

CRIDES stands for Centre de Recherche Interdisciplinaire Jean Renauld "Droit, Entreprise et Société". Several members of the research center investigate issues related to artificial intelligence from a legal perspective, in the various area of law represented within the center. Their academic approach is completed by their practical experience but also by insights gained through dialogues with various stakeholders.

The present contribution is a commentary of the Assessment List published by the High-Level Expert Group on Artificial Intelligence as part of the Ethics Guidelines for Trustworthy AI.

Remark #1

Depending on the exact purpose of the list, the level of detail is either excessive or insufficient

Remark #2

Some issues cristalyze differently depending on the stage at which the assessment is performed and related questions thus cannot be applied uniformily at all stages

Remark#3

The relation between some requirements included in the list and the existing legal requirements is to a certain extent unclear: Are they going beyond? Are they a mere restatement thereof? Are they only partially overlapping therewith?

Remark #4

The list is said to be grounded in the fundamental rights of the EU Charter and ECHR but its exact interaction therewith is unclear.

Remark #5

The list and its current level of detail do not take into account the complex mix of competences necessary to a comprehensive implementation thereof. There is no alternative offered to entities who wish to implement it but have limited resources.

Remark #6

The current phrasing leaves to much interpretative leeway and in some cases, this may undermine the quality of the performed assessment

[1] Phd researcher at the Law Faculty of UCLouvain (CRIDES).

[2] Professor of Law, UCLouvain (CRIDES) and USL-B, KULeuven, Munich Intellectual Property Law Center.

REMARK #1

Depending on the exact purpose of the list (to offer a toolbox for all situations or a fine-tuned instrument for assessing a particular AI application), the level of detail is either excessive or insufficient. Indeed, explanations included in the "Ethics Guidelines for Trustworthy AI" ('Ethics Guidelines") and preceding the assessment list specify as follow :

- **Purpose 1 : To offer a toolbox for all situations.**
  The list claims to build a "horizontal foundation to achieve trustworthy AI" but beyond that, "different situations raise different challenges". Therefore, this foundation must subsequently be adapted to each specific case.
  - ⇨ However the list appears too detailed to merely constitute such horizontal foundation (*see subsequent remarks*).
  - ⇨ In addition, the value-added of this specific"horizontal approach" is not clear. It indeed adds to a landscape of more than 80 AI Ethical Guidelines which have been already developed and publicized, and which "all include the similar principles on transparency, equality/non-discrimination, accountability and safety"[3]. The risk of redundancy is high.
- **Purpose 2: To offer a fine-tuned instrument for assessing a particular AI application**
  The list is "intended to be a method to operationalize the commitment to rendering AI trustworthy (…), it is guidance for practitioners". It may thus be understood as an agenda whose role is to ensure that discussions about AI tools do cover all fundamental features of trustworthiness. It is specified that the list is "non-exhaustive" (as it is "not about ticking boxes but about continuously identifying and implementing requirements") but at the same time it is depicted as "concrete".
  - ⇨ Despite this presentation, the agenda appears not detailed enough to ensure satisfactory operationalization (*see subsequent remarks*).

Consequently, the existing draft could stand some cuts (especially on the broad principles already stated in other existing guidelines) while at the same time benefit from some additional details ensuring the announced concrete character.

Proposed solution

From a practical standpoint, a desirable solution to these shortcomings could be the creation of two complementary documents:

- for preliminary discussions, a short assessment list (1 or 2-page format), to be considered by the management when taking a decision about the development or deployment of an AI tool and attributing tasks on the basis thereof. This would remedy the fact the list is somewhat too detailed.
- for subsequent implementation, an explanatory leaflet clarifying various terms and elements thereof, to be read and used by people in charge of performing the various aspects of the assessment. The latter could include more precise guidelines depending on the sectors and

[3] See https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/ (accessed Oct. 27, 2019).

sensitivity of issues. This would in turn remedy the fact the list is not detailed enough to operationalize the underlying principles.

REMARK #2

On top of a list of issues to be checked (but not as a list of boxes to tick), whether at a preliminary high level or at subsequent more detailed level, it appears important to design a simple process for selecting the questions relevant depending on the stage at which the assessment is performed. Indeed, different issues are to be discussed for the *development* (including the design) of a new AI tool, for its *deployment* and for its *use*.The Ethics Guidelines do distinguish those three stages. The first one involves mainly the management and the engineers of companies developing AI solutions, while the second is geared at the organisations and the public sector which intend to deploy and offer the AI tools developed by others. Some issues crystalize differently depending on the concerned stage and not all questions can thus be applied uniformily throughout all the stages.

- ⇨ If we however associate each of the existing questions on the list with the stages it applies to, it is unveiled how some stages may be underassessed.
- ⇨ Notably, the moment when the decision about the possible AI deployment is discussed and taken is of great importance. It is located at the very outset of the deployment stage. It is important to ensure that all the right questions are addressed at that moment so that the best decision is taken *before* the proper deployment and use (and before people are affected by the use). The assessment list should serve to enhance the capacity of the organisations and public authorities to also assess when deciding *not* to use AI tools is the best outcome for the final recipients of the technology.
- ⇨ In addition, the guidelines do not refer sufficiently to the categories of more vulnerable populations, such as children, or of "minority" or "less-represented"[4] groups[5]. Some questions may need to be further specified when AI tools are applied to their situation and a data gap increases the risk of bias or discrimination[6].

Proposed solution

To remedy this shortcoming and render the list truly operational, it is necessary to match questions included on the list with the stages at which they are to be asked. This should appear clearly in the short version of the assessment list (see Remark 1). If a question applies to several stages, it is further important to specify the practical differences it entails in the assessment. This could be done in the explanatory leaflet.

[4] See the gender data gap analysed by C. Criado Perez, *Invisible Women. Exposing bias in a world designed by men*, London, Chatto & Windus, 2019.

5 B. Nonnecke, "AI, Human Dignity, & Inclusive Societies: Priority Recommendations to Better Ensure AI Doesn't Deepen Disparities for Vulnerable Populations & Minority Groups", Medium, 30 May 2019.

[6] A recent case involves the Apple Pay tool for assessing creditworthiness which appears to favor men over women: https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html.

In line therewith, it is necessary to better indicate in the process for implementing the assessment list that a decision *not* to deploy or *not* to use the AI technology can be adopted as the best solution, and when and after what checks this decision should be taken.

Finally, more regard should be given to the impact of the use on the most vulnerable recipients of the technology. This refocus could be made throughout the guidelines and nuances specified in the explanatory leaflet.

REMARK #3

The relation between some requirements included in the list and existing legal requirements is to a certain extent unclear: Are they going beyond? Are they a mere restatement thereof? Are they only partially overlapping therewith?

Explanations included in the guidelines indicate that although there are 3 aspects of trustworthiness, namely (1) legality, (2) ethics and (3) robustness, the guidelines do not cover legality. Their aim is to specify things which go beyond the legal aspect. It is however acknowledged that the law does to a certain extent tackle elements ensuring ethicality and robustness, therefore creating possible overlaps. For the remaining, they rely on the assumption all legal requirements are met beforehand.

- ⇨ The unclear relation between the legality aspect on one side, and ethics and robustness on the other side, creates potential confusion and vagueness. Notably, at least two main headings in the assessment list (#2 Technical robustness and safety and #3 Privacy and data governance) are interlinked with legal obligations in the field of security/safety and in the field of data protection. It is not clear how on those issues, the assessment list offers additional safeguards to those contained in the law.
- ⇨ The list should also mention that the assumption that all legal requirements are met is not fully defendable. Several recent cases demonstrate that complying with legal requirements is not always easy as their correct implementation may in some cases remain difficult to assess. Accordingly, non-compliance may result merely from lack of understanding and foresight.

Proposed solution

Correlations between the existing legal requirements and the requirements included in the assessment list should be clearly specified. For the questions related to legal requirements which are usually difficult to implement, it is necessary to include guidance on how to comply therewith before specifying things to be done "beyond".

Accordingly, we notably propose to reconsider the two headings #2 and #3 as the overlap with the existing legal obligations is too broad. These sections could be used to clarify the legal framework should the latter be unclear, or complete it. In both cases, it is necessary to avoid needless and potentially confusing overlaps.

REMARK #4

The Ehics Guidelines highlight that the guidelines themselves and the resulting assessment list are "based on an approach founded on fundamental rights". We agree these should follow closely the framework for good behavior and decisions contained in the fundamental rights umbrella of the EU Charter and ECHR.

- ⇨ Yet the assessment list does not clearly distinguish where it translates the requirements of fundamental rights and where it goes further in regulating the development and deployment of AI tools.This is for instance the case with the headings #4, #5 and #7 about transparency, non-discrimination and accountability which are not clearly derived from the fundamental rights framework although they undeniably are related thereto.
- ⇨ The assessment list also contains a subsection devoted to fundamental rights in general but then includes the aforementioned sections #4, #5 and #7 .Moreover, it seems odd to include fundamental rights as a subsection in the section devoted to "human agency and oversight": human rights are a much broader topic as the subsequent mention of privacy and non-discrimination under separate sections do suggest.

Proposed solution

For the sake of clarity, we suggest to better structure the list:

- first by distinguishing the issues that prolong the fundamental right requirements and the other ones which are based on a sound regulation of technology;
- second, by structuring the fundamental rights-based issues *per fundamental right* (taking into account that the privacy concerns are already largely taken into account in the legislation on data protection).

REMARK #5

The list and its current level of detail do not take into account the complex mix of competences necessary to a comprehensive implementation thereof. There is no alternative offered to entities who wish to implement it but have limited resources.

Explanations included in the guidelines expressly highlight the risk that a lack of diversity of skills and competences may exist within entities supposed to perform the assessments. Indeed, if the guidelines are addressed to "all AI stakeholders designing, developing, deploying, implementing, using or being affected by AI, including but not limited to companies, researchers, public services, government agencies, institutions, civil society organizations, individuals, works and consumers", the assessment list itself is in particular addressed to developers and deployers of AI, whether internalized or acquired/licenced from third parties. Therefore it is acknowledged that it might be necessary to involve other stakeholders inside or outside the organization.

- ⇨ Such involvement may not necessarily always be possible.
- ⇨ It is thus necessary to render the list operational even for teams operating with limited skills. Especially within entities being SMEs, the lack of diversity of skills

may be more striking in respect to certain aspects of the suggested list while the small teams handle most tasks internally without the possibility to delegate tasks to external and more competent parties.

⇨ Furthermore, the assessment list does not clearly enough acknowledges that some points on the agenda may need to be handled not only by different units of a same entity but by different entities. Such is the case when AI is developed by a service company and then handed over to the company which will effectively use it on a daily basis.

Proposed solution

It may be desirable to

- specify for each point on the assessment list a capital question to be assessed in any case (even if the team is small) and phrased with enough clarity and detail as to allow anyone to perform an assessment based thereof.
- better specify key concepts. Indeed, within some entities, people involved in the assessment may have the technical skills but only a limited understanding of some other considerations. Notably the explanatory leaflet should include:
  - ⇨ A list of concerned fundamental rights and what aspects of AI may affect each of them;
  - ⇨ A list of possible bias (for instance, an indicative list of recurring bias). Considering people conducting the assessment may be at the origin of bias incorporated into the system, it must be kept in mind that it is a nearly impossible task to identify its own bias without knowing what precisely to look for.
- specify how the assessment list is to be divided should the creation, testing and use of the concerned AI system be executed by different entities.

REMARK #6

The current phrasing leaves to much interpretative leeway and in some cases, this may undermine the quality of the performed assessment.

⇨ Too many questions included in the list require a YES or NO answer while using adjectives such as "appropriate" and "suitable" or simply asking about the existence of a specific procedure.

Proposed solution

To properly assess what is appropriate or suitable on a case by case basis, some indications and examples of best practices could be welcome, to be included – for instance – in the explicative leaflet (see Remark 1). Indeed, if a procedure was set up within the concerned entity, someone certainly considered it as "suitable" in the first place. Chances are that the same person will end up being in charge of completing the assessment part related thereto and it would be seldom that this person suddenly criticizes her/his own work.

Besides, if verifying the very existence of procedures is a good starting point, it may be desirable to explicitly mention internal tracking of their use should be done. Notably, this could result in reports on events which proved those procedures to be fallible as well as on proactive updates and reasons thereof.

COMPLEMENTARY REMARK

Better understanding of the list and its concrete application in various situations may be gained over time. In this respect, the guidelines indicate that it is desirable to "disseminate results (of assessments) and open questions to the wider public". Experiences would thus ideally be shared. How is this envisaged in practice? For assessments based on the general requirements listed under Chapter 2, it is feasible. Yet for assessments based on the list specified under Chapter 3, competition (and trade secrets) considerations will undeniably prevent full dissemination of results. Indeed, the list is supposed to trigger discussions about what is clear but also what points raise doubts. If an entity uses AI in the course of its business, it will not openly disclose information about how the said AI raises interrogations based on questions included in the list. This would result in acknowledging that the entity generates a potential risk and accepts it on a "wait and see" basis.