AI4Belgium

Feedback Report on the "Trustworthy AI Assessment List"

in the context of the Piloting Process of the Ethics Guidelines for Trustworthy AI

by AI4Belgium

A. Introduction

AI4Belgium welcomes the opportunity given by the European Commission's High-Level Expert Group on AI (AI HLEG) to provide feedback on the "Trustworthy AI Assessment List" published in April 2019 as part of the Ethics Guidelines for Trustworthy AI.

AI4Belgium is a community-led initiative, enabling Belgian individuals and organisations to capture the opportunities of AI while facilitating the ongoing transition towards the technology's increased adoption in a responsible manner. AI4Belgium has the ambition to position Belgium firmly on the European AI landscape, drawing on the many assets vested in the Belgian AI ecosystem, from high quality researchers, excellent entrepreneurs and companies, to innovative public entities.

This feedback report was prepared in the context of an interactive workshop where the members of the AI4Belgium community came together and - with guidance of some experts of the AI HLEG - drafted a first feedback. This feedback was subsequently circulated within the AI4Belgium community, allowing all members an opportunity to consult the feedback document and provide their further input. The end result comprises the consolidated feedback of the AI4Belgium members.

The report first provides the key points of feedback raised by the community (B). Thereafter, suggestions are provided on how the Assessment List can be further improved, both in general terms (C) and regarding a number of specific topics (D).

B. Key points of feedback:

1. The publication of these guidelines is very useful, as it encourages people to reflect on the ethical issues that the development and deployment of AI-systems brings forth, and as it helps creating awareness around an important topic. Moreover, we welcome the fact that the Assessment List defines a common approach to deal with an issue that touches upon all levels of an organization.
2. We believe the Assessment List is too long in its current form. It should be significantly shortened and tailored to the specific context.
3. The procedure of the Assessment list is too burdensome for SMEs and start-ups, both in terms of the governance mechanism required by the list, and its actual use.
4. Many questions of the Assessment List are closed (and hence not very helpful) and/or unclear.
5. The Assessment List – and the requirements for its compliance – should become more concrete in case it were to move towards a regulatory framework.
6. It should be ensured that this Assessment List can keep evolving together with the technology.
7. A clarification is needed regarding who is considered the 'user' of an AI-system, and of the respective obligations / rights of the 'user'.

C. General suggestions for improvement:

- The main point of feedback raised, concerns the fact that it is currently unclear how the Assessment List can be used in practice. As the List is very long, the workload would be too high – in particular for smaller organisations – if it were to be used in every AI-related project. As a suggestion to remedy this concern, a more vertical approach which provides specific lists for different sectors, would be welcomed. Moreover, the use of a tree-like structure to go through the Assessment List, guiding users towards the relevant questions, evaluating early on if a certain branch requires an in depth assessment, and providing practical examples of how certain issues could be remedied, can likewise be encouraged.
- SMEs and start-ups may have a particularly hard time to implement the entire Assessment List in practice, both at the level of governance mechanism and actual application of the List. Ensuring that the use of the Assessment List does not disproportionally burden those entities is one of the main concerns that should be addressed. In this regard, the thoroughness of the assessment may need to vary in accordance with the risks posed, and with the size of the project undertaken.
- Some of the questions in the Assessment List seem to be overlapping or duplicated. It would be advisable to align and/or merge those questions, which will also help rendering the List shorter. In addition, some of the questions are rather vague and merit simplification, especially as they would need to be used and understood throughout the entire organisation, by people with different levels and areas of expertise. Furthermore, the questions do not always clarify what risk they are aiming to tackle, which could be helped by providing more practical examples.
- Guidance on the next steps after going through the Assessment List would be useful. For example, if there is a negative reply to a certain question, what can be done to improve the situation? What methods or solutions are available and advisable?
- Where possible, a prioritization of the topics and questions could also prove advantageous.
- Some definitions have been provided in the Guidelines' document, but there are still many words mentioned in the Assessment List which are not defined, and/or which have multiple definitions or meanings depending on the context. The AI HLEG has already provided an extensive document concerning the definition of AI, which could potentially be updated with additional definitions. For example, a detailed definition on "bias" and "explainablity "would give the document more clarity. Moreover, a number of concepts are too vague and/or broad (e.g. "unacceptable" risk), which would benefit from either some practical examples or a clearer scope indication.
- An additional section dedicated to conducting a "Risk Assessment" would be advisable. This could not only offer additional value to the Trustworthy AI Assessment List, but could also be integrated into other domains as a practical tool to be used by organisations. Many organisations already have standardised risk assessment procedures in place which have proven their usefulness, and this could be used as a basis for this purpose.
- As the guidelines are voluntary and often complex, a system could be set up where organisations choose their level of dedication (from full implementation to general considerations). This might improve the implementation rate for SMEs.
- Fundamental rights are the foundations of these guidelines, but in practice people with a technical background who work on AI will not have any practical experience with what those

rights entail. Additional guidance attached to these guidelines on the interplay of AI and fundamental rights could hence prove useful.

- Finally, a number of concerns were raised related to the legal approach which is currently being planned by the European Commission. A cautionary approach was mentioned, indicating that overly descriptive regulation could hinder progress in the development and deployment of AI. In turn, a regulatory sandboxing approach could prove suitable in this scenario.

D. Topical suggestions for improvement:

- On the topic of **human agency and oversight**: The questions raised have different levels of abstraction (from general autonomy-related questions to specific questions on chatbots), and it would be beneficial to ensure a similar abstraction level. Some questions seem to narrow, such as the question relating to AI-systems implemented in work processes which unnecessarily seems to exclude non-work processes. Moreover, as AI will potentially have a huge impact on the organisation process, there will be changes to the way organisations will function. More emphasis could be given to the fact that the implementation of AI will necessitate an analysis of workflows and tasks and the potential for new roles and responsibilities for impacted human operators, which in turn may necessitate a comprehensive management plan to ensure that the transition occurs as smoothly as possible. Finally, a number of the questions could be made more practical, by asking practitioners to for instance distinguishing the actions they take with the AI-systems' predictions (used only to assist? used in parallel to human decision-taking? used to replace?) and the objective reasons for this choice.
- On the topic of **privacy and data governance**: Many of the items mentioned under this section seem to be closely related to – and in some instances even overlapping with – the requirements set in the GDPR. Whilst the subject is undoubtedly of great importance, double work for organisations should be avoided, and the distinction between the Assessment List and the GDPR should be clarified. Questions regarding the impact of data quality on AI models should also be made more explicit.
- On the topic of **explainability**: The questions listed under this section do not always clearly specify to whom certain aspects should be explainable, which aspects should be explained, and to which extent this should happen. As people who have no knowledge of mathematics or computers are also stakeholders in this process, guidance on the extent and manner in which (the inner working of) the AI-system could be explained to them would be welcomed. Full explainability (for those stakeholders) can be a challenge, both in terms of feasibility and in terms of practicality, yet should be striven towards to the largest possible extent. In this regard, it would be useful to clarify the extent of the explanation that should be given to the various actors involved (including for instance government mandated inspectors in case of confidentiality or privacy issues). For non-technical stakeholders, providing an understanding of the true potential and limitations of the implemented AI-solution in order to form realistic expectations about the AI-system's performance would be important. Finally, one of the listed questions relates to the assessment of the system's business model, which does not seem to have a direct link with the system's explainability as such; a clarification in this regard would be welcomed.
- On the topic of **liability**: The document does not provide any guidance for organisations as to where their responsibilities begin and where they end. An AI-system typically exists as part of a whole value chain with many different actors involved. While a different Commission expert

group is currently focusing on issues regarding AI-liability, guidance on this topic would also be well-placed in the context of the Assessment List, and hence it would be useful to include this subject in the revised version to the extent possible.

- On the topic of **transparency**: Whilst increasing transparency is to be applauded, the practical implementation thereof seems to be difficult. Even if issues regarding the black box problem are left aside, difficulties may still arise regarding the required level of understanding of mathematics. A certain literacy on the subject seems required, since an oversimplification of explanations (which also related to explainability) could prove counterproductive. Further guidance on the required level and scope of transparency would therefore be useful. It is, for instance, not entirely clear whether transparency covers only the internal working of the implemented (learning) algorithm or also aspects like the nature, quality and volume of the input data used to train the data model and validate the algorithm. Additionally, since transparency on the data level is also a crucial step in achieving trustworthy AI, the inclusion of questions on "data authenticity" and how such authenticity can be verified could be added to cover this aspect in more detail.