Microsoft
Corporate, External & Legal Affairs
Rue Montoyer 51, B-1000 Brussels

30 November 2019

Introduction

Microsoft welcomes the opportunity to provide feedback on the Trustworthy AI Assessment List 2.0. developed by the European Commission's High-Level Expert Group on AI. From past efforts to prioritize trustworthy computing and digital privacy to an emphasis on Responsible AI today, the company has consistently sought to move to a people-centered approach to technology development and deployment.

The Microsoft team responding to the Assessment List consisted of Kalyan Ayloo (Office365 Intelligence), Ilke Kaya (Outlook), Jacquelyn Krones (Office of Responsible AI), Cornelia Kutterer (EU Public Policy), Michael Phillips (Lead attorney to the Office of Responsible AI), Luke Stark (Research), Frank Torres (Office of Responsible AI) and Hanna Wallach (Research). The team had a variety of comments on the current draft, which it felt overall was a valuable and productive document.

Though this response focuses on suggestions for potential improvement, Microsoft congratulates the Commission for its work thus far. Microsoft has found the engagement process with the Assessment List tremendously useful in ensuring the company's own internal work on AI standards, principles and practices are in alignment with those of the Commission, and we are excited to share expertise and research with the Commission going forward.

General Comments

To answer almost all the Assessment List's questions, the team felt they would need to collaborate with others in the company to help ensure a full and effective response. However, the taxonomy of roles and responsibilities listed in the response survey often seemed limited, and did not reflect the full range of potential groups within the company whose expertise would need to be drawn on to adequately address any given question. The team recommends the Assessment List and related documents begin with an assumption that all questions will require a collaborative response from multiple members of an organization.

The team felt some of the sections of the Assessment List, and the questions within them, were not well sequenced. For instance, some sections began with broad questions on abstract themes, which would be daunting to most AI practitioners. The team recommends mapping questions more closely to the AI development and deployment lifecycle instead of thematically.

The team also strongly recommended diverse, concrete, and contrasting examples from different domains of AI development be incorporated into the Assessment List. Without these examples, the team noted the Assessment List's questions risked seeming both overly abstract and misunderstood as inapplicable to a wide range of AI developers. In addition, the team noted Assessment should specify when its questions are directed at AI developers, designers, or deployers (as there is considerable difference between each of these groups' needs). Such clarifications will be useful.

The consistency of wording across different sections of the Assessment varied widely. At some points there seemed to be different conceptual assumptions motivating the varying definitions, or at least a lack of precise alignment among the various authors. Making the Assessment List conceptually and definitionally consistent is critical for its clarity of use. In addition, too many of the Assessment List's questions were framed as merely requiring either "yes" or "no" answers, which neither provided granular insights into existing process, nor satisfy a robust oversight function. The team noted ensuring the Assessment List and any subsidiary documents have a workflow that allows for appropriate pause points to enable easy implementation is also key to its success.

More broadly, the team was at times confused by the overall goal of the Assessment. For example, in some instances wondered whether to prompt responsible, necessary thinking around open-ended questions regarding the impacts of AI systems, or to function as a more prescriptive regulatory template. Determining which aspects of the Assessment should be enshrined in regulation and other compulsory mechanisms, and which should be supported by other best practices, is a critical question for the Commission going forward. Moreover, some of the sections of the Assessment, such as universal accessibility, are applicable to a wide range of technologies. The team recommended focusing tightly on AI-specific questions within the broader framework of technology regulation, while also exploring how specific sectoral regulatory frameworks, such as those around financial services, might serve as models for AI oversight.

The team believes there is a strong role for the Commission to support enterprises with more limited resources to engage internally with these questions. This can be done through the provision of resources, such as links to relevant related definitions and regulations, b the creation of an EU expert team able to work with companies to improve their practices, and in the funding of research programs around user participation in AI systems, human-machine interaction, and the sociotechnical impacts of AI.

Domain-Specific Comments

Human Agency and Oversight

Fundamental Rights

The team agreed identifying and documenting tradeoffs at the level of fundamental rights is valuable as an exercise to situation technologists The team was concerned the structure and placement of this section within the Assessment was lacking: beginning with a broad assessment of rights impacts would be challenging for most engineering teams without further scaffolding around how these broad principles should be operationalized.

In regard to Q2.2, the team found the question challengingly broad as worded: algorithmic decisions are ubiquitous, and without a discussion of the appropriate threshold for transparency, such a guideline is hard to operationalize. Team members described the problem as a "delicate balance around over-warning vs. under-warning," and noted the importance of HCI/UX research to understanding and establishing notification thresholds.

Human Agency

In regard to Q4.2, the team noted a lack of clarity around the meaning of "safeguards," which could range from either an on/off switch for a particular feature or even an entire system, or more granular and varied mechanisms to assess a system's impact. Once again, the team noted the importance of HCI/UX research to understanding/addressing this challenge.

Human Oversight

The team observed a lack of clarity regarding what group the questions in this section assume has human control over an AI system: is it the end user, the operator, or the editor/developer? As a result, the team noted there is a need for different versions of the Assessment List for system developers vs. deployers/end users.

Technical Robustness and Safety

Resilience to Attack and Security

The team was concerned with the wording of Q8, noting that it is both overly broad and tautological: it is not possible to *a priori* verify a system's performance under unexpected conditions if those conditions are indeed unexpected.

Fallback plan and general safety

The team was unsure at what level Q10 sought planning: at the level of imminent or immediate physical harm to users, or longer-term secondary harms to society more broadly. Both are important concerns but entail sacrificing different kinds of expected benefits to minimize potential harms. Regarding Q13, the team noted developers require clear metrics and KPIs (key performance indicators) to estimate harms in actionable ways.

Accuracy
In general, the team found this section useful and most of the questions well crafted. However, the team had concerns regarding the narrowness of how, and in what ways, "accuracy" was defined: focusing on technical definitions of fairness in statistical classification, which seemed implicit in the questions, omits a much broader questions around how to judge accuracy across a wide range of AI systems. As such, the scope of these questions needs to be broadened and their wording changed to reflect a wider array of AI domains.

Reliability and reproducibility

The team noted with concern that, the results of both personalized and probabilistic decision-making systems would not be reproducible in the terms laid out by this section's questions. Further, the team felt the questions in the section were also overly narrow. Given the broad range of domains and systems in which AI is being deployed, the team noted the exact method for supporting the requirement will depend on the particular system and where/how it is deployed. The team recommended focusing this section instead on whether developers can trace, log, and keep track of the system's activities at different points in its analysis.

Privacy and data governance

Respect for privacy and data protection
Regarding Q19, the team was concerned the question is unclear regarding to whom mechanisms to flag privacy protection issues should be aimed towards internal stakeholders or end users.

Quality and integrity of data
The team was impressed by this set of questions in terms of their wording and scope. The team noted "standards" as described in Q25 can mean a wide variety of metrics and was thus somewhat unclear.

Access to data
In regard to Q29.3, the team noted organization SLTs (senior leadership teams) must be

committed both to ethical principles and to building out support for practices/infrastructure to support them within organizations. The latter task is much more challenging for smaller companies, and it is an open question how to support this Assessment List in firms with limited resources.

Transparency

Traceability
The team had major concerns with the layout and framing of the questions in this section. The team suggests that a better way to ask such questions would be to break each down to address specific components of the system (e.g. data sets for training and testing).The team noted Microsoft has invested considerable resources into developing models for traceability in AI systems, which the company would be happy to share with the Commission.

Explainability
The team found this section extremely uneven. Regarding Q34.1, the team observed that managing explanations to end users in the process of engaging with AI systems was largely an HCI/UX problem. Moreover, the team flagged the distinction, which seemed missing in these questions, between informing users about a decision and justifying that decision to users: it is not clear when and how each type of explanation would apply to various parts of an AI system or experience.
This difference should be clarified.

Regarding Q34.2 and Q.34.4, the team found the questions intriguing, but extremely broad and requiring expertise in organizational anthropology/business management more than technical AI development. Q36 seemed high-level and out of place within the order of the rest of the section; Q36.2 seemed more appropriate for the following domain around bias and fairness, and the team found Q36.3 poorly worded and redundant. . The team thought it worth flagging that eyes-off data is sometimes a necessary constraint to explainability, for instance to preserve privacy in the case of email data. The presence of such “values trade-offs” is not well addressed in the Assessment List itself, though it is noted in the Commission’s other materials.

Communication
The team was not clear on what differentiates this segment from that on explainability, as the questions had similar themes. In general, the team found these questions challenging, and too much divorced from examples of actual systems. Broadly, the problems described in these questions are also HCI/UX design ones.

Diversity, non-discrimination and fairness

Unfair bias avoidance
The questions in this section generally led to productive conversations, but the team felt they were sometimes in the wrong order: for instance, the team felt Q44 should be the first question of the segment, and Q44.3 should also be higher in the question order. The team found Q43 similarly problematic to Q8 above: both are unclear especially the meaning of the word "conditions" and tautological.

More broadly, the team observed that high-level guidelines are not enough to ensure items in the Assessment List are operationalized. Considerable work is needed to develop checklists and other forward guidance for practitioners on the ground. Such guidelines require some amount of specialization based on domain, organization, and team. Thus, one overarching regulatory document will likely not provide either practitioners or regulators adequate flexibility.

Accessibility and universal design
The team found this section's and the following section's division of questions confusing. Q45 and sub-questions around accessibility are a prerequisite for all technical development (at least in the United States, which has relatively robust laws around accessibility)—but AI developers do not handle these elements of design. Q46 and sub-questions seemed better suited to the stakeholder participation section.

Stakeholder participation
As noted, the team found these questions overlapped significantly with Q46 and sub-questions; the two sections should either be consolidated or properly divided. The team notes stakeholder participation research is vital but often under-resourced.

Societal and environmental well being

Sustainable and environmentally friendly AI
The team expressed a high degree of concern and urgency around issues of sustainability and the climate emergency. The team noted that both the general nature of the problem, and the structural mismatch between those concerns and what can be done with current computing infrastructure, made the questions in this section difficult to answer.

Social Impact/Society and Democracy
The team found these questions stimulating, though Q53 seemed overbroad. The team

noted that it is not always within the ability of lower level employees in an organization to make decisions regarding business directions and broader social effects of a product (such as social disruption and job loss). The extent to which individual employees have agency and responsibility for such decisions is an open question.

Accountability

Auditability
The team noted the questions in this section overlapped heavily with those addressing *Traceability* section above, and the section was thus redundant.

Minimizing and reporting negative impact
The team found the definition of risk in this section to be vague, and noted there is little guidance around how product teams should engage with legal frameworks or external guidance being developed or already in place.

Documenting trade-offs
The team found this section useful and its questions well worded.

Ability to redress
The team noted the questions in this section overlapped with those in the *Transparency* section and were somewhat redundant.

Conclusion

Microsoft appreciates the opportunity to provide these comments and suggestions. The company is currently in the process of implanting its own internal policies around Responsible AI (https://www.microsoft.com/en-us/AI/our-approach-to-ai), and seeks to support its customers in developing and deploying AI responsibly in a variety of domains. We look forward to providing whatever assistance we can to the Commission in strengthening its Assessment List and other governance mechanisms for ensuring responsible and trustworthy AI.