Olle Häggström
Professor of mathematical statistics
Chalmers University of Technology

November 29, 2019

SUBJECT: Evaluation of the Pilot "Assessment List" for Trustworthy AI

Dear EU HLEG Members and the EU Commission,

We greatly appreciate the opportunity to evaluate and comment on the "Assessment List" for Trustworthy Artificial Intelligence (AI) released by the European Union's High-Level Expert Group (HLEG) on AI. We also look forward to future opportunities to engage with the EU HLEG and the Commission on the ideas presented below, as well as other pertinent topics as a "European Approach to AI" is developed.

Summary Feedback

To begin, we commend the EU HLEG for developing an Assessment List of this kind. In doing so, the HLEG has taken on the extremely difficult task of trying to operationalise a theoretical, ethical framework into a useable format for AI developers and deployers. We believe strongly that this type of Assessment List can help establish real, ethical practices for all *AI actors*.[1] Though the Assessment List is "primarily addressed to developers and deployers of AI systems,"[2] we also hope that consumers (e.g., government agencies) and end-users of AI systems will use the list to help inform their own thinking about the types of questions that they will want to have answers to *before* establishing how much to trust particular AI systems. While highly supportive of this effort, we do have two high-level comments for the consideration of the HLEG and Commission. In addition, we have also provided numerous concrete edits and suggestions to the list itself in the second portion of this document.

First, beyond the content of the Assessment List itself, it is extremely important for the HLEG and Commission to evaluate the potential market and legal benefits that may be provided to AI actors that comply with this Assessment List. As noted already by the

[1] As defined by the Organisation for Economic Co-operation and Development (OECD) in the agreed upon "Recommendation of the Council on Artificial Intelligence," *AI actors* are "those who play an active role in the AI system lifecycle, including organisations and individuals that deploy or operate AI." See: OECD, "Recommendation of the Council on Artificial Intelligence," Section I, at https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.
[2] High-Level Expert Group on Artificial Intelligence, "Ethical Guidelines for Trustworthy AI," p. 24, at https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

DEPARTMENT OF MATHEMATICAL SCIENCES
Division of applied mathematics and statistics
Chalmers University of Technology
SE- 412 96 Gothenburg, Sweden
+46 31-772 5311
olleh@chalmers.se
https://www.chalmers.se/en/Staff/Pages/olle-haggstrom.aspx

Chalmers University of Technology
Corporate Identity No: 556479-5598

CHALMERS
UNIVERSITY OF TECHNOLOGY

HLEG, we agree that compliance with the Assessment List by AI actors should *not* be considered "evidence of legal compliance, nor is it intended as guidance to ensure compliance with applicable laws."[3] Perhaps more importantly, we also believe that compliance with the Assessment List should *not* grant AI actors or their AI systems limited or blanket liability protection by various EU or EU Member States' legal systems. Components of this Assessment List are too subjective for compliance with it to afford any amount of *explicit* liability protection by governments, e.g., through regulations or other legal mandates. For example, please review the questions provided by Section 2, on "Technical robustness and safety," under the subheading "Resilience to attack and security."[4] Two different AI developers that carefully follow the Assessment List, both in good faith, could develop vastly different subjective answers to questions such as what it means to *ensure* "the integrity and resilience of the AI system against potential attacks" or what *suitable* preventative measures are for dual-use technologies. Therefore, the end value of having completed the Assessment List, and the amount of liability the developers and deployers retain for use of their AI systems, can only be evaluated on a case-by-case basis *ex-post* by various legal systems.

Other experts, such as the Data Ethics Commission of the German government (Datenethikkommission), have also noted the importance of carefully assigning liability for AI systems. In the context of this Assessment List, we second their recommendation that "if harm is caused by autonomous technology used in a way functionally equivalent to the employment of human auxiliaries, the operator's liability for making use of the technology should correspond to the otherwise existing vicarious liability regime of a principal for such auxiliaries."[5] Further, we believe that the retention of liability by AI actors, despite the use of the Assessment List, is in accordance with the Accountability principle of the OECD, of which many EU Member States are already signatories.[6]

Despite not thinking the use of the Assessment List should convey explicit liability protection, we encourage the HLEG and Commission to evaluate ways to incentivize the use of the Assessment List by AI actors in other ways. For example, the HLEG and Commission may wish to consider ways in which use of the Assessment List by AI actors can be formally recognized to convey a potential market advantage for the AI system amongst consumers (e.g., through branding with a certified label). As another example, use of the Assessment List could also be one of many conditions required to be considered for procurement by EU or EU Member State

---

[3] Ibid, p. 26.
[4] Ibid, p. 27.
[5] German Data Ethics Commission, "Opinion of the Data Ethics Commission: Executive Summary," (official English translation), Recommendation 74, p. 26, at https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.pdf?__blob=publicationFile&v=2.
[6] The "Accountability" principle states that "AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art." Retention of liability by all AI actors encourages each to carefully monitor and maintain proper functioning of AI systems. See: OECD, "Recommendation of the Council on Artificial Intelligence," Section IV, 1.5, at https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

agencies. Regardless of the method, we think the Assessment List is an important tool that should be used by AI actors, just not one where it reduces the liability--and therefore the legal accountability for outcomes--of AI actors.

Second, to maximize the benefit of the Assessment List, the HLEG and Commission should expand on the guidance it provides to developers and deployers as to *when* in the process of development or deployment the tasks prompted by the questions can and should be completed. Some of the Assessment List's tasks, prompted by its questions, can be relatively easily understood as to *when* in the process of the development and deployment they should be completed.[7] Others, however, are more difficult, and could have differential effects if "completed" at different points in time by developers and deployers. For an obvious example, one of the Assessment List's questions is "Did you verify how your system behaves in unexpected situations and environments?"[8] Hypothetically, this verification could be done by a well-meaning AI developer in an "pilot" or "soft launch" phase of deployment, where the AI system is being used for a small sample of real-world use cases. Another AI developer, however, may make a greater effort to verify system behavior through simulation or through carefully controlled experiments, *not* with real-world use cases. The latter approach may be harder to do, but is important for establishing the trustworthiness of higher consequence AI systems. While we acknowledge that the HLEG has stated that this Assessment List "will need to be tailored to the specific use case and context in which the system operates,"[9] we believe some further guidance on how to tailor the timing of these tasks is especially important for AI systems meant to be used for critical societal functions.

Specific Comments and Suggestions for the Assessment List

In addition to the summary comments above, we also have specific suggestions and comments on the content of the Assessment List. To ease possible review, we have provided these in an itemized fashion below, with some rationale for the suggestion/comment provided following the suggestion or comment. As a reference, when we provide new text for consideration, we do so in *underlined italics*, and eliminated text is shown with ~~strikethrough~~. When we provide a comment on existing text, we [bracket] the language we are commenting on, followed by our comment.

[7] For example, the tasks prompted by the questions "Did you communicate to (end-)users – through a disclaimer or any other means – that they are interacting with an AI system and not with another human? Did you label your AI system as such?" are obviously meant to be done after the creation of the AI system but before their deployment and use. See the section on Transparency, subheading Communication, in High-Level Expert Group on Artificial Intelligence, "Ethical Guidelines for Trustworthy AI," p. 29, at https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
[8] Ibid, section on Technical robustness and safety, subheading Resilience to attack and security, p. 27.
[9] Ibid, p. 24.

Comments and suggestions in "1. Human agency and oversight"

Fundamental rights:

- Did you carry out a [fundamental rights] impact assessment where there could be a negative impact on fundamental rights? Did you identify and document potential trade-offs made between the different principles and rights?
    - **Comment**: It may be difficult for AI developers and deployers to identify all of the "fundamental rights" that are necessary to be assessed. We recommend either providing an example list of fundamental rights to consider, or identifying an existing list, such as the Universal Declaration of Human Rights.

- Does the AI system interact with decisions by human (end) users (e.g. recommended actions or decisions to take, presenting of options)?
    - Could the AI system affect human autonomy by interfering with the (end) user's decision-making process in an unintended, *harmful* way?
        - Rationale for addition: Not all unintended effects to human autonomy are harmful. For example, a school teacher may start using a new interactive, AI learning system to help guide the introduction of new elements of a math curriculum into the classroom. In doing so, they may *intentionally* (by design of the system) spend less time "autonomously" planning on how and when to introduce new math material to the class. However, *unintentionally,* the teacher may also: A) spend less time and effort autonomously reviewing the class's understanding of the new material, assuming that the new intelligent learning system's recommendations are appropriately matching their classroom's capabilities, and/or B) spend more time providing enriching activities in art and other subject areas. These *unintentional* effects are not equal in their effects, where A) is more likely to be harmful, and B) is likely to be innocuous or beneficial.

Human oversight:

- Did you consider the appropriate level of human control for the particular AI system and use case?
    - Can you describe the level of human control or involvement?
    - *Did you evaluate* who ~~is~~ *are* the "human*s* in control" *and what roles they have in the development or deployment of the AI system,* and what are the moments or tools for human intervention?
        - Rationale for addition: It is important for developers and deployers to consider "humans in control" as a continuum, so that all AI actors have accountability over the functioning of an AI system. This is in accordance

with the aforementioned Accountability principle of the OECD.[10] For example, a CEO of a multinational corporation deploying an AI system may decide that all hiring managers *must* use a particular AI system to make their hiring decisions, and so in a relative sense, that CEO is "in control" of the use of the AI system. However, many more humans would be in control and have accountability for its proper functioning if each hiring manager had the *option* to use the system. Further, even if they *must* use the system, the hiring managers in the company would be in control of *how* the system used on a day-to-day basis.

Comments and suggestions in "2. Technical robustness and safety"

Fallback plan and general safety:

- *Did you use any explicit validation methodologies?*
    - *Did you put any technique in place to make any exploration the system does as safe as necessary?*
    - *Did you create mechanisms to mitigate foreseeable negative side effects of operation?*
        - Rationale: It is important for AI developers to identify and use explicit validation methods that are being formulated by leading AI safety researchers, especially to address the problem of "safe exploration" in AI systems and to mitigate known side effects.

Accuracy:

- *Did you ensure that the level of accuracy of the system to be expected by users is properly communicated?*
    - *Have you communicated a segmentation of cases where you expect it to be more accurate and less accurate?*
        - Rationale for addition: Both developers and deployers should communicate clearly to users the expected level of accuracy (i.e., error rate) of a particular AI system, especially if it has varying error rates for different types of use cases. For example, with computer vision systems, the error rates for proper identification of objects may be higher in stable, indoor lighting, than in outdoor settings with variable lighting.

Reliability and reproducibility:

- *Did you develop a method for the system to calculate and report a confidence score in each recommendation? Did you evaluate the confidence score to ensure it is well-calibrated?*

[10] The "Accountability" principle states that "AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art." Retention of liability by all AI actors encourages each to carefully monitor and maintain proper functioning of AI systems. See: OECD, "Recommendation of the Council on Artificial Intelligence," Section IV, 1.5, at https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

- *Did you develop a method for the system to detect if it is being used in an online usage setting that reflects a materially different statistical distribution than that on which it was trained? If so, does the system use that information to lower its confidence score and report it to the user?*
    - Rationale for additions: If an AI system can and does communicate a confidence score, it reinforces both the trust in the accuracy of the AI system and allows the user to better understand factors that may make it less reliable than others. This is particularly important if the AI system is being deployed for use cases it was not trained extensively for by the developers. This is even true when the reported confidence score is low, as it allows the human user to appreciate the benefits and limitations of the particular system in given use cases.

- *Did you develop a mechanism to evaluate when the AI system has been changed enough to merit a new review of its technical robustness and safety?*
    - Rationale for addition: AI systems are not static, and must be reevaluated by their developers and deployers for continued technical robustness and safety as they evolve. This can be due to continued learning from the data the AI system receives while deployed (including possibly "poisoned" data), or external software updates or changes in the use pattern of the AI system by the deployers.

We also recommend creating a new subheading called "Safety Culture" under 2. Technical robustness and safety after the existing subheadings.

Safety Culture:

- *Did you provide proper training in the latest AI research on technical robustness and safety to your employed AI developers or deployers so they can identify possible problems?*
- *Did you create an easy mechanism for individuals concerned about the technical robustness and safety of your AI system to report it to the proper authorities, or to make the appropriate fixes to the system themselves? Are individuals informed about how to raise concerns through this mechanism?*
- *Did you create a mechanism to incentivize improvements to the technical robustness and safety of developed or deployed AI system?*
- *Did you foresee any kind of external, governmental guidance for technical robustness and safety applicable to your AI system, and incorporate those in addition to internal initiatives?*
    - Rationale for addition: One of the major ways in which both developers and deployers can create trustworthy AI systems is to nurture a robust safety culture within their staff. These and other questions like it will help make sure those responsible for developing and deploying the AI system have the appropriate training and authority to improve upon the technical robustness and safety of AI systems. The last question, "Did you forsee…" is a variation of a question that appears later in the Assessment List on "Accountability"

Comments and suggestions in "4. Transparency"

We recommend creating a new subheading called "Fiduciary responsibility" under 4. Transparency, in between the subheadings for "Explainability" and "Communication."

Fiduciary responsibility:

- *Did you evaluate and communicate in whose interest is the AI system primarily designed to operate: the end-user, the developer, or another intermediary deployer?*
    - *If the AI system is designed to benefit someone other than the end-user, is this transparent to the user? Are any conflicts-of-interest disclosed?*
    - *How are you modeling the interests and goals of the users (and other relevant persons or organizations) so as to understand conflicts between them?*
- Rationale: A key component of trust in an agent is knowing who exactly it is working for. Many AI systems can benefit both the end-user and the developer/deployer of the system, but in ways that are not transparent to either (most commonly, it is opaque to the end-user). For example, recommendation systems for media often benefit the end-user by providing "better" recommendations on which media to consume next, but those same systems may also benefit the developer/deployer by capturing preference data or even information about how to shape future preferences of other users. Likewise, a user should know when a "recommendation" is having its weight increased by a paid advertiser for the recommended product in addition to the user's past preferences. As AI agents and assistants become more powerful and widespread conflict-of-interest are likely to do the same.

Communication:

- Did you clarify the purpose of the AI system and who or what may benefit from the product/service?
    - *Did you communicate segments of usage scenarios where the system is expected to perform better and worse? Did you specify the specific types of biases the system has or deviations from the norm for each of these?*
        - Rationale for addition: As noted with comments to the "Accuracy" subheading previously, it is important for end-users to understand in which use cases the system is able to perform better or worse.

Comments and Suggestions for "5. Diversity, non-discrimination and fairness"

Unfair bias avoidance:

- Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design?
    - *Did you consult with social scientists regarding the proper tradeoffs to choose among any residual unwanted biases which cannot be eliminated?*

        - Rationale: AI developers and deployers are unlikely to have the expertise of many social scientists needed to evaluate, if necessary, the harm caused by particular biases that cannot be eliminated. When tradeoffs are possible, such as by increasing AI systems tolerances for either false-positive or false-negative results that reflect different biases, social scientists should be consulted to determine a possible best path forward.

Comments and Suggestions for "6. Societal and environmental well-being"

Social impact:

- In case the AI system interacts directly with humans:
    - *Did you design or deploy the system with an explicit or implicit goal of it being an addictive or attention-maximizing system?*
    - *Did you include warnings to users if you suspect the AI system may stimulate addictive behavior in users?*
        - Rationale: There is growing concern that AI systems can be used to "hack" human behavior in many ways, to include increasing the addictive use of particular systems to benefit the interest of developers/deployers. Therefore, warning the user of this effect is essential to trustworthiness.

Society and democracy:

- *Did you assess whether the AI system disproportionately facilitates or automates antisocial, authoritarian, violent, or disinformational activities? If the system facilitates or automate such social ills, what mechanisms did you create to mitigate this?*
    - Rationale: As with many dual-use or omni-use technologies, it is important for developers to consider whether an AI system is more likely to be used for negative purposes, *especially* as it might undermine society and democracy as a whole. For example, a developer is currently weighing whether and how to release an AI system capable of accurate lip-reading given its many potential misuses as well as benefits for those that are hard of hearing.

Background on Authors

Dr. Olle Häggström is a professor of mathematical statistics at Chalmers University of Technology, a researcher at the Institute for Future Studies in Stockholm, and a member of the Royal Swedish Academy of Sciences. He is the author of 90+ scientific papers and four books, most recently "Here Be Dragons: Science, Technology and the Future of Humanity."

Dr. Max Tegmark is a professor doing physics and AI research at MIT, and advocates for positive use of technology as President of the Future of Life Institute. He is the author of over 250 publications as well as the New York Times bestsellers "Life 3.0: Being Human in the Age of Artificial Intelligence" and "Our Mathematical Universe: My Quest for the Ultimate Nature of Reality." His AI research focuses on intelligible intelligence.