Comments of the Center for Democracy & Technology on European Commission's High Level Expert Group on Artificial Intelligence (AI HLEG)'s Draft Ethics Guidelines for Trustworthy AI

Introduction: Rationale and Foresight of the Guidelines

The Center for Democracy & Technology supports the High-Level Expert Group (HLEG)'s efforts to develop guidelines for trustworthy AI and appreciates the opportunity to comment on this draft. In particular, we commend the group for affirming a rights-based approach to governing AI, for moving beyond the development of principles, and for acknowledging the need for a context- and domain-specific implementation of the values discussed in these guidelines. While we agree that trustworthiness is a key objective for any system, the HLEG must also acknowledge the limitations of current methods for mitigating bias in machine learning models. In many contexts and applications, truly trustworthy AI remains hypothetical. Moreover, trustworthiness depends not only on the ethical purpose and technical robustness of the model or application but also on the governance of the entire societal context or legal system within which an AI application sits (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3265913). We recommend that the HLEG place greater emphasis on (1) the importance of mechanisms and processes for continually interrogating and challenging AI systems from both the inside and the outside and (2) the importance of assessing the entire system (including underlying policies, laws, and human-technology interactions) that surround the AI.

The draft guidelines avoid the pitfall of relying on ethics alone as a solution to mitigate the potential harms of AI systems. Many industry actors have embraced ethical codes, but they rely too heavily on the ability of a particular company, or individual or team within the company, to weigh the ethics of a system and decide what is best for everyone it will affect. As AI Now wrote In its 2018 report (https://ainowinstitute.org/AI_Now_2018_Report.pdf), "Ethical approaches in industry explicitly ask the public simply take corporations at their word when they say they will guide their conduct in ethical ways," and "ethical codes may deflect criticism by acknowledging that problems exist, without ceding any power to regulate or transform the way technology is developed and applied." Ethics may also fail to address how an open-source technology may be used by others who are not bound by a particular ethical code. Instead, the HLEG is right to point out that the development, deployment, and use of AI must "respect fundamental rights and applicable regulation."

Similarly, we support the HLEG's effort to move beyond principles to more concrete guidance. However, as the HLEG acknowledges, general high-level guidance cannot address the context- and domain-specific challenges, ethical considerations, or rights implications of an AI application. The European Commission is setting an important example for the rest of the world by affirming that competitiveness in AI need not and should not come at the expense of human rights or ethics.

While we support the goal of trustworthiness, we caution that for many AI models that have been or are currently being developed, the trustworthiness of the model itself remains

hypothetical. Research into methods for removing or avoiding harmful biases in machine learning models is progressing (http://proceedings.mlr.press/v81/), but we are still far from solving these problems (https://www.omidyar.com/insights/public-scrutiny-automated-decisions-early-lessons-and-emerging-methods). Thus, rather than "trust in technology," the current moment calls for an emphasis on trustworthiness in the systems within which the technology is used and the processes that govern its use.

Chapter I: Respecting Fundamental Rights, Principles and Values - Ethical Purpose

Once again, we commend the HLEG for re-affirming the EU's "rights' based approach to AI ethics." However, when fundamental human rights are translated to ethical "principles and values" to govern AI, it is likely that different stakeholders and decision makers will apply the principles differently. This is particularly true of beneficence ("do good"). We have found that between civil society and industry, and even among industry actors, beliefs about what technologies or designs benefit society can diverge widely.

Moreover, in many applications, it may be far from clear how the values articulated by the HLEG should be balanced against each other. For example, one focus on AI fairness/ethics research in the EU and the US is on how to create recommender systems (for news, entertainment, trending topics, jobs, etc.) that are more equitable with respect to the diversity of publishers that are able to reach an audience or that promote a more diverse (or less polarized) information diet (https://piret.gitlab.io/fatrec2018/program/). This work advances values such as non-discrimination and beneficence (they arguably "do good" by reducing polarization or disinformation in news dissemination). However, it could also be seen as interfering with human autonomy by nudging people toward content that they wouldn't otherwise choose or suppressing the effects of majority preferences. It is likely that different stakeholders would balance these values very differently.

We recommend that HLEG include more discussion of the right to an effective remedy (or redress) in this section. Specifically, "explicability" should serve not only to inform citizens about the existence and operation of AI systems to build trust but also to facilitate effective appeal and remedies. This is particularly important given that, no matter how much due diligence is performed, AI systems will continue to make mistakes.

Chapter II: Realising Trustworthy AI

Governance:

The guidelines discuss the need for governance of both data and "AI autonomy," but they are missing a discussion of governance that extends to the larger system or context within which the AI is deployed or with which it interacts. For example, automated decision systems are being developed and deployed to replace or (more often) assist with human decision making.

When researchers and civil society study the ethics or fairness of these systems, we often find that the problems are not necessarily (or not only) within the AI or even the deployment and governance of the AI but deeper within the pre-existing system or structure. No matter how technically robust or ethically designed an AI system is, it will not be trustworthy if it is designed to execute decisions in a system that itself is untrustworthy.

Privacy:

As the HLEG acknowledges, it is critical to test AI systems' performance on different subgroups, particularly vulnerable and minority groups, in order to identify and mitigate discrimination. This may require the collection or inference and use of sensitive characteristics. Collecting or inferring this information while maintaining appropriate privacy protections raises challenges without easy answers, and this will be a critical area for legal and technical analysis over the next few years. Privacy laws are critical but should not become a barrier to assessing AI for discrimination and protecting vulnerable groups. The Commission should consider providing guidance, with input from affected communities, on ways to collect sensitive-characteristic test data while complying with the GDPR.

Documentation and ethical constraints in open AI:

Machine learning models and training data sets are often made available for anyone to use and incorporate into their own project or product, or train using their own dataset. In order to ensure values such as safety, non-discrimination, and robustness, it is not enough for the original developer or data collector to hold themselves to those values. They must consider how their designs or datasets may be used and iterated on by others, including malicious actors. Researchers have come up with ideas for ensuring that open models (https://arxiv.org/abs/1810.03993) and open data (https://arxiv.org/abs/1803.09010) are not misused, including models for documenting the robustness and domain of the model/data and methods for putting fairness constraints into the code (https://dl.acm.org/citation.cfm?id=3287588). However, it will take a multipronged approach and continuous auditing to ensure that open models are not used in untrustworthy ways.

Traceability and auditability:

These are necessary characteristics for trustworthy AI. However, the draft guidelines over-emphasize the function of transparency for facilitating "laypersons'" understanding of "the causality of the algorithmic decision-making process and how it is implemented . . ." and " the laymen's acceptance of the technology." First, the "layperson" distinction may be misplaced, since even someone with deep expertise in machine learning will not inherently "understand" how every model works and will benefit from auditability. Second, even if useful and meaningful explanations of AI systems can be developed, the people who are affected by those systems should not assume the burden of truly understanding how they work, or how a particular automated decision causes a particular effect on a person's life. For the average person, the

choice to an engage with an AI system in some way is unlikely to be truly optional. Thus, “acceptance” of the system should not be presumed from the existence of an accessible explanation of the system. Instead, a primary goal of transparency should be to facilitate appeals and redress when an AI system does something wrong.

Stakeholder engagement:

The draft guidelines correctly identify the need to ensure the participation and inclusion of stakeholders in the design and development of AI systems that will impact them. However, the guidelines do not elaborate on potential ways to do this. Stakeholder participation is a value that is often identified but rarely operationalized. HLEG could add value by developing concrete recommendations for engaging stakeholders, including the adoption of processes, such as hearings and town halls, where community members can weigh in on AI applications being considered for deployment in their communities.

Chapter III: Assessing Trustworthy AI

The draft proposal includes many helpful questions for the assessment of Trustworth AI. In addition to those, please consider including the following questions:

Design for all:
Do people have a non-AI alternative or substitute for the system or service? Considering the availability of alternatives, or lack thereof, may inform the degree to which users of an AI system do so out of need, versus choice.

Safety:
Have the effects of the risk mitigation or management plan been tested? From a trust perspective, proof of the effectiveness of risk mitigation and management is an important metric.

Has there been an assessment of the potential to improve the risk mitigation or management measures? Where such measures have been tested, trust in systems could be enhanced by evidence of learning from past iterations.

Have the effects of interactions between multiple AI systems been identified? As AI systems become more ubiquitous, they may be more likely to interact with each other and produce unanticipated effects. Attempts to at least identify the potential risks associated with such interactions could be helpful in an assessment of trustworthiness.

Transparency:
Purpose-

Is/are the system(s) being used as intended? As systems are deployed and used, it will be important to understand how, in what contexts, and for what purposes they are actually used.

What measures have been taken to limit unintended uses? Where unintended uses are known, or anticipated, measures designed to limit either the use or the effects of unintended uses could help to ensure that fundamental rights are not inadvertently infringed.

What impacts might the system have on the fundamental rights of the intended users? Given the rights-based approach for assessing Trustworthy AI, an impact assessment through the lens of fundamental rights should be included.

What impacts might the system have outside of the intended group? Since AI systems may be able to impact non-users, an impact assessment for the potential of a system to affect their fundamental rights should also be part of the overall assessment framework.

When systems make decisions impacting people other than the user, such as in autonomous driving systems, are the criteria for balancing risks and benefits to the public communicated? The general public should have a means of accessing information regarding how AI systems will measure and weigh the risks imposed through their use.