COMMENTS OF PUBLIC KNOWLEDGE

Date: February 2, 2024
Re: NIST Docket No: 231218-0309, *Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)*

Public Knowledge thanks the National Institute of Standards and Technology for its December 21, 2023 Request For Information, and the opportunity to assist NIST in carrying out its responsibilities under the Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence issued on October 30, 2023.

Developing Guidelines, Standards, and Best Practices for AI Safety and Security.

Public Knowledge supports the development of robust, empirically-backed guidelines, standards, and best practices that can be used to enhance the overall trustworthiness of AI tools and systems. Accountability mechanisms such as certifications, audits, and assessments serve a crucial purpose in establishing a trustworthy and socially beneficial AI ecosystem.

Developing consensus-based industry standards around accountability mechanisms with multi-stakeholder participation and public review is an important stepping stone towards a more comprehensive regulatory framework for managing and mitigating AI risk. These standards should be viewed as potential tools which will likely need to be deployed alongside other regulatory instruments to build a truly trustworthy and publicly beneficial AI ecosystem. It is important to recognize that these procedures alone cannot ensure safe and accountable AI, and may need to evolve flexibly as the technology and AI marketplace continues to evolve and develop. Ultimately, voluntary standards will need to be supported by an enforceable regulatory framework, drawing upon existing consumer, product, and public safety style regulations—ideally, managed by an independent and expert federal regulatory agency.

At the same time, even in this early phase of standards setting, we caution NIST to pay careful attention to the needs and concerns of academic researchers, AI developers working on open models, and smaller AI companies. The AI ecosystem is hurtling towards consolidation and oligopoly, with the largest tech companies already dominating the field. The largest and most well-resourced actors must be held to the highest standards but should not be permitted to dominate the standards setting process in order to hedge out competitors.

Developing a Generative AI Companion Resource for the AI RMF.

As the RFI recognizes, generative AI presents new challenges and risks compared to decision-making and evaluative systems. AI tools that can create and manipulate images and videos, identify and expand on complex patterns of information, and operate in the field of language, present vast new frontiers of opportunity as well as danger.

Generative AI tools could be leveraged for the purposes of consumer manipulation, impersonation and deception, and enable never before seen scalability in the production of mis- and disinformation. And risks are not limited to misuse: it has already been observed that the power and accessibility of tools that operate in natural language may foster overreliance and overconfidence despite inaccuracy and ineffectiveness. Finally, many of the issues remain the same as well. Generative AI systems are vulnerable to bias, raise concerns about privacy, and the specter of the displacement or denigration of human labor in the face of automation looms over every AI innovation.

These are the risks observed and imagined now, as innovation and development is proceeding at a rapid pace; there will undoubtedly be new and unimagined challenges in the future as well. As a result, it is particularly encouraging that the RFI inquires about the diversity of "professions, skills, and disciplinary expertise organizations need to effectively govern generative AI" and what role they have to play in ensuring a comprehensive perspective on risk, safety, and accountability.

An interdisciplinary approach across a variety of academic areas including philosophy, ethics, social sciences, and cultural studies must be integrated with technical and engineering perspectives. Industry and academic researchers must be consulted along with a range of civil society organizations to offer perspectives on economics, public policy, and civil and human rights.

Forms of transparency and documentation.

Transparency and documentation are key tools that will be effective for communicating important information—and enabling accountability—for developers, deployers, and end users. Model cards, system cards, benchmarking results, and impact assessments can be valuable when developers are building off of other models; can help deployers make informed decisions about features and functionality of models as fit to their purposes;

and users can use this information to choose products and evaluate the reliability of outputs.

One especially important component of transparency is training data. Data set transparency can serve an important accountability function in enabling third-party assessment and evaluation. However, we are concerned that requiring AI developers to publish reports, summaries, and evaluations of training datasets could serve as a compliance pitfall for open-source, non-commercial, or competitive AI developers with more limited resources and team. Public disclosure of the full training data used should be the goal, and should satisfy any transparency or disclosure requirements. Simple full disclosure of the full training datasets is the ideal for enabling maximum transparency and enabling third-party evaluation anyway, so standards should ensure that compliance is maximally transparent while as simple and straightforward as possible.

Creating guidance and benchmarks for evaluating and auditing AI capabilities.

We are encouraged by the role NIST is taking in the process of developing benchmarks and assessments of AI risks and capabilities. As we have previously written, "[e]ffective government oversight of AI systems will involve active participation in the writing and development of assessments. By shaping assessment criteria, the government prevents private entities from hijacking accountability measures and ensuring that evaluations remain aligned with the public interest. This approach will lead to robust standards that address transparency, privacy, security, and fairness, enabling comprehensive evaluations and continuous improvement. Government involvement in assessment development also fosters democracy and inclusivity by ensuring diverse stakeholder input and enhancing public trust in the oversight process"[1]

Ultimately, we advise that: "Oversight and accountability must be coupled with policies and investments that will promote continued innovation, responsible research, open access, and vigorous competition. AI technologies have enormous potential, and the overall goal of any regulatory system should be realizing that potential, not overburdening its development or allowing its benefits to be inequitably captured."[2]

Reducing the Risk of Synthetic Content.

One of the key challenges posed to democracy by AI systems is their ability to further distort the integrity of our information environment. More (and increasingly credible) disinformation will lead to continued declines in citizens' trust in news and other

[1] https://publicknowledge.org/policy/ntia-ai-accountability/
[2] https://publicknowledge.org/policy/ntia-ai-accountability/

democratic institutions, as well as having the potential to create harm and spark violence. Disinformation narratives, whether of domestic origin or foreign, also prevent people—including policymakers—from solving our most pressing problems.

Generative AI systems can compound the challenges in our information environment in at least three ways: increasing the number of parties that can create disinformation narratives, making them less expensive to create, and making them more difficult to detect. Traditional cues that alert researchers to false information, like language and syntax issues and cultural gaffes in foreign intelligence operations, will be missing. This isn't just about AI "hallucinations" – researchers have already proven that clean, convincing news articles, essays and television scripts can be purposefully created using AI. Image generators, may undermine the classic entreaty to "believe your own eyes" to determine what is true and what is not.

Technical measures has so far proven mperfect and may be outpaced by developments in the technology itself. It seems unlikely these tools will win a technological arms race with motivated generators of disinformation.

Therefore, it is critical to focus on robust information ecosystem scale solutions. There is no silver bullet for dealing with disinformation generated by malicious actors, but a combination of trusted and reliable fact-based journalism institutions, user-facing tools for transparency and information, AI and information literacy and education, robust content moderation standards, and real oversight and accountability for the digital platform and AI sectors, would all contribute towards developing a healthier information environment with users that are more resistant to the dangers of malicious disinformation.

* * *

We thank you for the opportunity to comment on these pertinent issues and look forward to further opportunities to engage with NIST on ensuring a safe and accountable set of standards for AI accountability.