March 18, 2024

The Honorable Gina Raimondo
Secretary
U.S. Department of Commerce
1401 Constitution Ave NW
Washington, DC 20230

via email

Re: Ensuring "AI safety" into the future begins with how we address algorithmic harms now

Dear Secretary Raimondo,

We, the undersigned civil society, tech policy, workers' rights, consumer protection, science advocacy, civil liberties, and racial justice organizations—including organizations that are participating as members in the U.S. AI Safety Institute (AISI or Institute) Consortium—write to articulate our shared expectation that the National Institute of Standards and Technology (NIST) continue to foreground a broad view of "AI safety," one that accounts for the entire range of algorithmic harms.

Complementing the work of enforcement agencies and civil rights offices to prevent and remediate algorithmic discrimination, NIST has played a critical role in shaping new standards and inviting multi-stakeholder dialogue on actionable AI governance. While we recognize that efforts to govern AI warrant some attention to novel risks that may be posed by certain systems, this work should not come at the expense of efforts to address AI's existing impacts that threaten people's opportunities, freedoms, and right to a healthy environment.[1]

Importantly, building a foundation for future safeguards demands focusing attention on the demonstrated real-world harms affecting people now. Methodologies to identify, measure, prevent, and remediate today's harms are evergreen, enabling NIST to build the long-term governance muscle ultimately needed to mitigate novel emerging risks.

Put simply: Addressing the theoretical risks of AI begins with addressing the ways AI is harming people now.

[1] *See*, *e.g.*, Olga Akselrod, *How Artificial Intelligence Can Deepen Racial and Economic Inequalities*, ACLU (July 13, 2021), https://www.aclu.org/news/privacy-technology/how-artificial-intelligence-can-deepen-racial-and-economic-inequities; Christina Swarns, *When Artificial Intelligence Gets It Wrong*, Innocence Project (Sept. 19, 2023), https://innocenceproject.org/when-artificial-intelligence-gets-it-wrong/; Alexandra Sasha Luccioni, et al., *Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model*, 24 Journal of Machine Learning Research 1 (June 2023), https://www.jmlr.org/papers/volume24/23-0069/23-0069.pdf; Merve Hickok & Marc Rotenberg, *The UK AI Summit: Time to Elevate Democratic Values*, Council on Foreign Relations (Sept. 27, 2023), https://www.cfr.org/blog/uk-ai-summit-time-elevate-democratic-values ("The AI safety agenda should not ignore the AI fairness agenda.").

1. **NIST's attention to more speculative AI harms should not compromise its track record of scientific integrity.**

Due to NIST's central role in the federal government's efforts to advance safe and trustworthy AI, many are looking to NIST's leadership to set the example on managing AI risks, measuring harms, and protecting the American public. **That is why it is all the more critical for NIST to move forward with its long-held commitment to sound measurement science across the full range of AI harms.**

We understand that some provisions of Executive Order 14110 specifically directed NIST to pay attention to emerging novel risks,[2] and we appreciate that NIST has sought to quickly onboard experts in those particular domains. However, the Executive Order also charges NIST with important work to support implementation of the minimum practices for safety- and rights-impacting uses of AI by federal agencies; to create guidance and benchmarks for evaluating and auditing AI capabilities more broadly; to create guidance for effective AI red-teaming; and to engage in the development of global AI standards, among other tasks that extend beyond novel risks of emerging AI models. NIST must ensure its staffing and resourcing efforts meet all of these needs.

While the evidence on AI's existential risks remains in many ways speculative,[3] a large body of evidence indicates that AI and algorithmic systems are producing serious and tangible harms to people now. Evidence demonstrates AI's negative impacts on workers' jobs and economic opportunity,[4] excessive use of scarce resources such as water and energy,[5] racially biased outcomes in medical treatment,[6] arbitrary decisions in social and medical benefits,[7] civil rights abuses in policing and the justice system,[8] and

[2] Executive Order 14110 of October 30, 2023, 88 Fed. Reg. 75191 (Nov. 1, 2023), § 4.1.

[3] Further, research on such risks is often unempirical. As detailed in a recent letter by the U.S. House of Representatives Committee on Science, Space, and Technology, research findings suggesting AI's world-ending risks are "often self-referential and lack the quality that comes from revision in response to critiques by subject matter experts." They also sometimes use discredited evaluation methods or do not undergo academic peer review. *Letter to Laurie Locascio*, U.S. House of Representatives Committee on Science, Space, and Technology (Dec. 14, 2023), https://republicans-science.house.gov/_cache/files/8/a/8a9f893d-858a-419f-9904-52163f22be71/191E586AF744B32E6831A248CD7F4D41.2023-12-14-aisi-scientific-merit-final-signed.pdf. (omitting internal citations).

[4] Alexandra Mateescu & Aiha Nguyen, *Explainer: Algorithmic Management in the Workplace*, Data & Society Research Institute (Feb. 6, 2019), https://datasociety.net/library/explainer-algorithmic-management-in-the-workplace/; Alexandra Mateescu, *Challenging Worker Datafication*, Data & Society Research Institute (Nov. 8, 2023), https://datasociety.net/library/challenging-worker-datafication/.

[5] Pengfei Li, et al., *Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models*, arxiv (April 6, 2023) https://arxiv.org/abs/2304.03271.

[6] Jesutofunmi A. Omiye, et al., *Large language models propagate race-based medicine*, npj Digital Medicine (Oct. 20, 2023), https://www.nature.com/articles/s41746-023-00939-z.

[7] Casey Ross & Bob Herman, *Denied by AI: How Medicare Advantage plans use algorithms to cut off care for seniors in need*, STAT (Mar. 13, 2023), https://www.statnews.com/2023/03/13/medicare-advantage-plans-denial-artificial-intelligence/.

[8] National Academies of Sciences, Engineering, and Medicine, *Facial Recognition Technology: Current Capabilities, Future Prospects, and Governance*, National Academies Press (2024), https://nap.nationalacademies.org/catalog/27397/facial-recognition-technology-current-capabilities-future-prospects-and-governance; Kashmir Hill, *Eight Months Pregnant and Arrested After False Facial Recognition Match*, New York Times (Aug. 6, 2023), https://www.nytimes.com/2023/08/06/business/facial-recognition-false-arrest.html.

military applications that may undermine human rights and constitutional values.[9] These areas pose urgent questions about rigorous, scientific methods to evaluate, audit, and address AI harms that require NIST's further engagement. At a time of divergent approaches to auditing and risk mitigation, NIST should advance sound, sociotechnical measurement across AI applications.[10]

NIST's continued close attention to these harms is necessary to ensure safe, secure, and trustworthy AI. While some progress has been made in developing approaches to address such harms, there is still a long way to go to consolidate these methodologies, test their efficacy, and incentivize their adoption. NIST and the administration should not leave this work unfinished.

2. The best way to approach the evolving set of risks posed by AI is to set evidence-based methodologies to identify, measure, and mitigate harms.

NIST's continued engagement on current harms need not detract from its additional mandate to explore potential harms of future AI capabilities. Indeed, NIST's actions to establish durable governance mechanisms are just as applicable to novel emerging risks as they are to those affecting people now. Principles in the AI Executive Order and in NIST's AI Risk Management Framework—such as pre-deployment testing, explainability, assessment of impact, ongoing measurement, and harm remediation—are evergreen across the range of algorithmic harms. They are useful not just to mitigate the risks of present day issues like algorithmic discrimination, but would also enable accountability regimes and human interventions to discover and safeguard against novel threats, such as the risk of AI hijacking critical infrastructure systems.[11]

Accordingly, NIST should ensure that the remit of its Safety Institute addresses not only unfamiliar risks from highly-advanced AI systems, but also concerns posed today by more prosaic automated decision systems.

Finally, across all harms under consideration, we expect to see NIST adopt a sociotechnical lens—one that understands the complex ways technologies interact with people and institutions, and the impacts those interactions have—to shape the whole of its risk management and safety toolkit.[12] For example, testing systems for technical vulnerabilities is an important component of mitigating cybersecurity risk,

---

[9] Katrina Manson, *US Used AI to Help Find Middle East Targets in Airstrikes*, Bloomberg (Feb. 26, 2024), https://www.bloomberg.com/news/articles/2024-02-26/us-says-it-used-ai-to-help-find-targets-it-hit-in-iraq-syria-and-yemen; Arthur Holland Michel, *Is AI the Right Sword for Democracy?*, Just Security (Nov. 13, 2023), https://www.justsecurity.org/90067/is-ai-the-right-sword-for-democracy/.

[10] Amy Winecoff & Miranda Bogen, Trustworthy AI Needs Trustworthy Measurements, Center for Democracy & Technology (March 6, 2024), https://cdt.org/insights/trustworthy-ai-needs-trustworthy-measurements/.

[11] *See* Matt Goerzen, et al., *Entanglements and Exploits: Sociotechnical Security as an Analytic Framework*, 9th USENIX Workshop on Free and Open Communications on the Internet (FOCI '19) (2019), https://www.usenix.org/conference/foci19/presentation/goerzen (proposing a framework of sociotechnical security against a multitude of threats and vulnerabilities).

[12] *See*, *e.g.*, Laura Weidinger, et al., *Sociotechnical Safety Evaluation of Generative AI Systems*, Google DeepMind (Oct. 18, 2023), https://arxiv.org/abs/2310.11986; Madeleine Clare Elish & Elizabeth Anne Watkins, *Repairing Innovation: A Study of Integrating AI in Clinical Care*, Data & Society Research Institute (Sept. 30, 2020), https://datasociety.net/library/repairing-innovation/; Andrew Selbst, et al., *Fairness and Abstraction in Sociotechnical Systems*, FAT '19: Proceedings of the Conference on Fairness, Accountability, and Transparency (Jan. 2019), https://dl.acm.org/doi/10.1145/3287560.3287598.

but understanding the human behavior and institutional norms surrounding those systems can be equally—if not more—important to securing the nation's safety from cyberattacks.

Conclusion

With the unwelcome news of budget cuts at NIST,[13] we can appreciate that hard choices may need to be made about the Institute's immediate priorities. In that difficult context, it is critical that the Institute continue to advance NIST's long-standing commitment to scientific integrity in service of the American public.

We commend the many ways NIST has focused on the need to address a wide range of harms that people already face from AI and algorithmic systems, and to apply sociotechnical methods to addressing those harms. We urge you to ensure that the Executive Order's additional focus on emerging AI concerns does not overshadow the many well-known harms that warrant continued attention and investment.

For any questions or further discussion, please contact Brian J. Chen (Policy Director, Data & Society) at brianc@datasociety.net or Miranda Bogen (Director of the AI Governance Lab, Center for Democracy & Technology) at mbogen@cdt.org.

Respectfully,

Data & Society
Center for Democracy & Technology
Accountable Tech
AFL-CIO Technology Institute
AI Now Institute
AI Risk and Vulnerability Alliance
American Civil Liberties Union
Center for AI and Digital Policy
Center on Race and Digital Justice
Communications Workers of America
Consumer Reports
Electronic Privacy Information Center
Fight for the Future
Government Information Watch
GovTrack.us
Kapor Center
Mozilla
National Employment Law Project
New America's Open Technology Institute
Public Citizen
Union of Concerned Scientists
Upturn

[13] Cat Zakrzewski, *This agency is tasked with keeping AI safe. Its offices are crumbling*, The Washington Post (Mar. 6, 2024), https://www.washingtonpost.com/technology/2024/03/06/nist-ai-safety-lab-decaying/.

Cc: Laurie Locascio, Under Secretary of Commerce for Standards and Technology
Elizabeth Kelly, Director of the U.S. AI Safety Institute
Deirdre Mulligan, Principal Deputy U.S. Chief Technology Officer
Bruce Reed, Assistant to the President and White House Deputy Chief of Staff
Ben Buchanan, White House Special Advisor on AI