February 2, 2024

Elham Tabassi
National Institute of Standards and Technology
100 Bureau Drive, Stop 200
Gaithersburg, MD 20899

Dear Ms. Tabassi,

On behalf of the Center for Data Innovation (datainnovation.org), we are pleased to submit this response to the National Institute of Standards and Technology's (NIST) request for comments to assist in carrying out its responsibilities under sections 4.1, 4.5 and 11 of Executive Order (EO) 14110 on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (AI).[1]

The Center for Data Innovation studies the intersection of data, technology, and public policy. With staff in Washington, London, Ottawa, and Brussels, the Center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as open data, artificial intelligence, and the Internet of Things. The Center is part of the Information Technology and Innovation Foundation (ITIF), a nonprofit, nonpartisan think tank.

In this submission, we make three main points:

1. NIST should work with other federal agencies to craft well-defined policy objectives for safety and trust in their domains to guide NIST's technical AI standards;
2. NIST should consider pre-generation attack vectors when exploring techniques to reduce the potential risks from AI-enabled synthetic content;
3. NIST should advocate for enhanced U.S. representation at international standards fora and coordinate better with the nation's global network of standards attachés to advance responsible global technical standards for AI.

[1] "NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence)," Federal Register December 21, 2023, https://www.federalregister.gov/documents/2023/12/21/2023-28232/request-for-information-rfi-related-to-nists-assignments-under-sections-41-45-and-11-of-the.

Hodan Omaar
Senior Policy Analyst
ITIF's Center for Data Innovation
homaar@datainnovation.org

Aswin Prabhakar
Policy Analyst
ITIF's Center for Data Innovation
aprabhakar@datainnovation.org

Nigel Cory
Associate Director, Trade Policy
Information Technology and Innovation Foundation
ncory@itif.org

1. NIST should work with other federal agencies to craft well-defined policy objectives for safety and trust in their domains to guide NIST's technical AI standards

Section 4.1(a)(ii) directs NIST to establish guidelines and benchmarks for evaluating AI capabilities that could cause harm.[2] As part of its assignment, NIST has specifically said it will develop guidelines for AI red-teaming and establish environments for test, evaluation, verification, and validation (TEVV) of AI systems' safety and trustworthiness.[3] The TEVV framework comes from the defense context; the Department of Defense (DOD) uses it for reviewing AI systems intended to be deployed in military applications, so Commerce creating a TEVV framework seeks to expand this tool for AI accountability to nondefense applications.

There are a number of problems with simply expanding TEVV to nondefense applications. The first problem is that the scope of application areas that Commerce's framework is supposed to cover is incredibly broad. Even if it only covers critical-impact AI systems, such as those that operate or manage critical infrastructure, are deployed in the criminal justice system, or used in a way that poses a significant risk to civil rights or safety, the TEVV framework will have to work for largely disparate contexts. NIST will have to create technical standards that are simultaneously able to robustly test an AI system that a judge wants to use to predict recidivism risk, an AI system that an energy company wants to use to optimize energy from the grid, and an AI system an edtech company wants to use to personalize learning in classrooms, to name three examples of infinitely many.

Moreover, the objective of this scheme is not clear. Accountability and transparency are means to an end, not an end in themselves, though some may say the end goal in all these contexts is to build "trust" with consumers. However, while "building trust" may fly as a public policy goal, creating technical standards forces NIST to bring analytic clarity to such vague goals. In the context of criminal justice, trust might mean systems are fair and unbiased; in critical infrastructure, trust might mean systems are reliable and safe; while in the context of education, trust might involve ensuring student privacy. The crucial term here is "might"—these are complex questions that many other agencies need to answer before the NIST can develop technical standards.

While NIST may be able to create a high-level approach to a TEVV framework for AI safety and reliability, it will need specific policy goals for different sectors to create more meaningful standards. To address this gap, NIST should host a workshop for invited federal agencies to develop domain-specific AI safety goals for key critical-impact AI systems. NIST should then use these policy goals to inform its work on developing standards for AI safety.

---

[2] "Test, Evaluation & Red-Teaming," *NIST* (December 2023), https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence/test.
[3] Ibid.

2. NIST should consider pre-generation attack vectors when exploring techniques to reduce the potential risks from AI-enabled synthetic content

Section 4.5 (a) of the EO directs the Secretary of Commerce to submit a report to the Director of the Office of Management and Budget (OMB) and the Assistant to the President for National Security Affairs identifying existing standards, tools, methods, and practices, along with a description of the potential development of further science-backed standards and techniques for reducing the risk of synthetic content from AI technologies.

The scope of NIST's report should cover the entire lifecycle of digital content. Currently, the EO focuses on potential content authentication protocols that kick in after the content, such as an image or text, has been generated. For instance, the EO wants NIST to explore post-generation content authentication methods such as tracking provenance; labeling synthetic content, such as using watermarking; and detecting synthetic content. But bad actors have several vectors of attack, some of which happen pre-generation. For one, they may attack the interface between hardware and software. Many devices have firmware that acts as a bridge between hardware and software and malicious actors could exploit firmware vulnerabilities in a camera for instance, to manipulate the raw data before it is processed into synthetic media. Others may inject malicious or misleading data into the training set used for training generative models to influence the behavior of the model, leading to the generation of synthetic media with unintended characteristics. Researchers from Harvard Medical School and the Massachusetts Institute of Technology have already explored how adversarial attacks that modify the pixels in medical images can manipulate them in ways that fool AI systems into misclassifying them.[4]

Given concerns about election misinformation and disinformation, the demand for a timely solution is understandable. However, NIST should not singularly focus on post-generation content authentication protocols because doing so may inadvertently leave gaps in the overall security framework. Pursuing a comprehensive strategy that considers both pre- and post-generation aspects will better position the agency to provide effective guidance in mitigating potential risks and address emerging challenges. Importantly, NIST has already done some of this work. For instance, NIST published a report last month identifying the types of attacks that manipulate the behavior of machine-learning (ML) systems.[5] As it notes, "The spectrum of effective attacks against ML is wide, rapidly evolving, and covers all phases of the ML lifecycle—from design and implementation to training, testing, and finally, to deployment in the real world. The nature and power of these attacks

[4] Scott G. Finlayson et al., "Adversarial Attacks on Medical Machine Learning," *Science* 363, no. 6433 (March 22, 2019): 1287–1289, https://doi.org/10.1126/science.aaw4399.
[5] Apostol Vassilev et al., "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations," *NIST Artificial Intelligence (AI) 100-2 E2023* (National Institute of Standards and Technology, January 4, 2024), https://doi.org/10.6028/NIST.AI.100-2e2023.

are different and can exploit not just vulnerabilities of the ML models but also weaknesses of the infrastructure in which the AI systems are deployed."[6] In its report to OMB, NIST should consider including the ways in which these types of cyberattacks specifically impact the creation of synthetic content. Similarly, NIST should explain how the Platform Firmware Guidelines it has published, which provide general guidance on various aspects of cybersecurity, specifically apply to the hardware and firmware components necessary to initialize AI systems for synthetic media.[7]

3. NIST should advocate for enhanced U.S. representation at international standards fora and coordinate better with the nation's global network of standards attachés to advance responsible global technical standards for AI

Section 11(b) of the EO directs the Secretary of Commerce to establish a plan for global engagement on promoting and developing AI consensus standards, cooperation, and coordination, ensuring that such efforts are guided by principles set out in the NIST AI Risk Management Framework and the U.S. Government National Standards Strategy for Critical and Emerging Technology.

On strategies for driving adoption and implementation of AI-related international standards, the U.S. government and like-minded nations should start by enhancing representation and influence in international standards of governance and leadership. The U.S. government's commendable efforts to get its candidate (Doreen Bogdan-Martin) elected as Secretary-General of the United Nation's International Telecommunications Union (ITU) is the clearest example of increased engagement and influence on technical standards in one standards development organization (SDO). The ITU deserves special attention as it does not have the same institutional governance arrangements (e.g., open, transparent, and consensus-based decision-making) as other standards bodies, which otherwise act as a safeguard against bad behavior and proposals from any country. The ITU is so often a focal point for debates about problematic Chinese technical standards proposals—like facial recognition—because it is a government-based body and states set the agenda as opposed to industry-led, multistakeholder SDOs where progress depends on consensus. However, it's unclear what else the United States has done since then. At recent ITU meetings, the in-person U.S. delegation was the same size as Saudi Arabia and smaller than Australia's.

NIST should also work with the International Trade Administration (ITA) to revise its standards attaché network. ITA employs standards attachés around the world who monitor emerging standards issues that have potential trade implications for U.S. industry and businesses. These entities are one of the main ways the U.S. government can directly engage in technical standards work and

[6] Ibid.
[7] Andrew Regenscheid, "Platform Firmware Resiliency Guidelines," *NIST Special Publication (SP) 800-193* (National Institute of Standards and Technology, May 4, 2018), https://doi.org/10.6028/NIST.SP.800-193.

discussions, but they are not even mentioned in the National Standards Strategy for Critical and Emerging Technologies the Biden administration published in 2023.[8] The U.S. government should revise the network so it has staff in the right countries tracking the most important standards-related developments, particularly for AI. There should also be clear reporting and integration with NIST because as ITIF explains in its 2023 piece, "Unpacking the Biden Administration's Strategy for Technical Standards: The Good, the Bad, and Ideas for Improvement," current reporting lines between the attaché network and the Department of Commerce are inefficient and misaligned.[9]

.

[8] Nigel Cory, "Unpacking the Biden Administration's Strategy for Technical Standards: The Good, the Bad, and Ideas for Improvement," (ITIF, October 2023), https://itif.org/publications/2023/10/10/unpacking-the-biden-administrations-strategy-for-technical-standards-the-good-the-bad-and-ideas-for-improvement.
[9] Ibid.