SUBMITTED VIA REGULATIONS.GOV

February 2, 2024

Information Technology Laboratory
ATTN: AI E.O. RFI Comments
National Institute of Standards and Technology
100 Bureau Drive, Mail Stop 8900
Gaithersburg, MD 20899-8900

RE: Request for Information on NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence

Dear NIST Information Technology Laboratory:

Intel Corporation ("Intel") appreciates the opportunity to provide comments to the National Institute of Standards and Technology ("NIST" or "Agency") in response to its request for information related to the Agency's responsibilities under the White House Executive Order ("E.O.") on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence issued on October 30, 2023. Specifically, the E.O. directs NIST to undertake an initiative for evaluating and auditing capabilities relating to artificial intelligence ("AI") technologies and to develop a variety of guidelines, including for conducting AI red-teaming tests to enable deployment of safe, secure, and trustworthy systems.

Intel plays an important role in AI. Intel's full spectrum of hardware and software platforms offer open and modular solutions which support AI workloads and fuel emerging usages like AI at the edge, pioneering innovations that will advance the future of AI and help to solve the world's most complex challenges. For example, in healthcare and life sciences, we accelerate research and patient outcomes with faster, more accurate analysis across precision medicine, medical imaging, lab automation, and more. For manufacturing, we transform data into insights that help companies optimize plant performance, minimize downtime, improve safety, and drive profitability. Intel is committed to advancing AI technology, including generative AI, responsibly and contributing to the development of principles, international standards, best practices methods, tools, and solutions to enable a more responsible, inclusive, and sustainable future.

We offer the following input on NIST's efforts related to generative AI risk management and other topics outlined in the Agency's notice.

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

Developing a Companion Resource to the AI Risk Management Framework ("AI RMF") for Generative AI

Just as the use cases and potential benefits of generative AI technology are broad and diverse, so are its implications and potential risks. For example, generative AI has been used to develop deepfake images and videos contributing to the spread of disinformation and to public distrust in the authenticity of certain online content. Additional risks may include "hallucinations" or inaccurate outputs, bias, security and privacy concerns, and copyright and legal risks. Generative AI may require thoughtful and meaningful guardrails to protect our society and economy. We support a risk-based, multi-stakeholder approach that leverages international standards (e.g., ISO/IEC) which can be adopted at scale and allow organizations to implement the most appropriate internal processes and policies that are suitable for the responsible development, deployment and use of their AI products and services. One key benefit of this approach is that it can evolve over time with generative AI innovations and advancements. We urge NIST to adopt such an approach in its development of a companion resource to the AI RMF for generative AI. We also encourage NIST to leverage existing AI risk management and generative AI-specific efforts including workstreams and policy recommendations from the President's Council of Advisors on Science and Technology's Generative AI Working Group for areas of overlap and opportunities for coordination. We recommend that NIST collaborate with its counterparts in other regions (such as the EU, Canada, UK, Japan etc.) to develop tools, methods, and benchmarks, leveraging existing international standards to the extent possible and contributing to new international standards as appropriate, as many of the AI E.O. assignments to NIST are of interest across countries.

*Roles That Can or Should be Played by Different AI Actors for Managing Risks and Harms of Generative AI (*e.g., *the Role of AI Developers vs. Deployers vs. End Users)*

In a risk-based framework, it is important for policymakers to understand the generative AI value chain and the respective roles of developers and deployers who are aware of the intended purpose of the technology. Generally, a risk-based approach outlines responsibilities for developers and deployers to identify and mitigate risks and specify the level of collaboration that may be required. As an example, a deployer would specify boundaries and evaluate the legitimacy of the intended use cases for its AI systems. An important consideration is that the organization that determines the purposes and means by which a generative AI system would be used has more visibility and thus should bear greatest responsibility. Training practices should be considered carefully. As initial training, testing or evaluation by a developer cannot possibly cover all foreseeable uses, some of these activities may also be incumbent on the deployer or user of the AI system.

As a best practice, accountability, transparency, and trust are important principles for organizations developing and deploying generative AI technologies. To help engender trust in generative AI systems, the public and end users should know that the technology is reliable, that bias mitigation is monitored and addressed, and that it includes tools to assess the authenticity of its outputs. AI developers should implement measures to facilitate these principles and assist in identifying content as AI-generated. AI developers can build safeguards within the technology to help prevent it from being used to generate disinformation, including detection tools which can provide automatic AI-generated content detection and marking. They can also provide, or recommend to an AI deployer, watermarking tools and other techniques to help make it clear to users when audio and visual content is AI-generated. It will be important to provide benchmarks and vetting processes to help assess the efficacy, security, and quality of these tools. In addition, a certified repository could be useful to reduce the barrier for access.

AI deployers should enable end users to be educated about the intended functions and limitations associated with its generative AI products and services. This includes close monitoring so that production datasets feeding the AI system are verified to be appropriate, relevant, and quality assured to operate properly and as designed and trained by the developer. In particular, deployers should monitor their production datasets for deviations, drifts, anomalies, and bias until retirement of the AI system to detect if and when re-training is needed, or if a new system needs to be deployed, or its operation ceased. They should also develop a multi-faceted approach to evaluate and test potential risks. This includes assembling diverse, multi-disciplinary teams to test risk mitigation strategies under a comprehensive array of scenarios that engage end users. Once all testing scenarios have been successfully navigated, the generative AI product could be fit for public release. Furthermore, AI deployers should establish and implement stringent data protection and security mechanisms to safeguard the release of personally identifiable information to third parties and prevent data breaches and unauthorized access.

Model Validation and Verification Including AI Red-Teaming

Comprehensive AI testing and red-teaming should focus on both AI safety (reducing potential harms) and security (defending systems against malicious activity). AI red-teaming includes the broader process of identifying AI vulnerabilities and failure modes for AI models and systems; we encourage NIST to establish guidance on red-teaming for both. AI red-teams are interdisciplinary technical teams comprised of experts such as, security researchers, privacy experts, prompt researchers, and model and data hackers whose activities are scoped to the level of risk associated with the AI system and/or its applications. Red-teaming should be conducted throughout the AI lifecycle (e.g., against sources of models and data, in-transit infrastructure, data prep stages, model repository, weights, parameters, vector databases, training infrastructure, inference stages, retrieval augmented generation (RAG) and response generation stages). The scope of red teaming should include traditional security threat analysis vectors but also extend to threats to unique AI usages including adversarial machine learning, poisoning, malicious/unintended biasing, malicious/unintended ethics violation, and malicious/unintended privacy violation. In addition to the AI model itself, focus areas should include the content (datasets) that they are trained on and content scanning for safety, security (malware), and privacy. Developers will need to implement controls earlier in the AI system development lifecycle denoting 'secure and safe' datasets.

Content Authentication, Provenance Tracking, and Synthetic Content Labeling and Detection

Training data and models should be protected at rest, in use, and in transit. Established confidential computing techniques can provide protection in use and combined with other existing techniques to provide protection at rest and in transit. Moreover, confidential computing techniques can cryptographically identify software by its measurement (cryptographic hash value or digest). Hence, a model and its surrounding software can be strongly identified and allow-listed, or equivalently, modification of the model and its software become tamper evident. Additionally, these techniques are auditable via remote attestation to collect hardware-endorsed evidence of the security state of the AI system.

Human Rights Impact Assessments, Ethical Assessments, and Other Tools for Identifying Impacts of Generative AI systems and Mitigations for Negative Impacts

As recognition grows of generative AI systems and their ability to pose societal risks such as privacy violations, bias, biometric surveillance, and disinformation campaigns, frameworks are needed to help organizations identify, prevent, and mitigate these potential harms. In its development of an AI RMF companion resource for generative AI systems, NIST should consider risks to human rights. In addition to other appropriate stakeholders, representatives from civil society organizations with human rights expertise can help to inform the development of risk assessment approaches that explicitly address human rights concerns resulting from generative AI systems. Proposed methodologies could be informed by international principles such as the

United Nations Guiding Principles on Business and Human Rights[1], Organization for Economic and Co-Operation and Development Guidelines for Multinational Enterprises,[2] and UNESCO Recommendation on the Ethics of Artificial Intelligence.[3] NIST could also consider providing guidance to organizations on how to assess potential downstream impacts on individuals and communities with greater risk of adverse human rights impacts due to their vulnerability or marginalization.

Reducing the Risk of Synthetic Content

Authenticating Content and Tracking Its Provenance

As video and animation have become more sophisticated, manipulated images can appear real and trustworthy, making the public more susceptible to deliberate deception and dis- and misinformation. Provenance and content authentication techniques such as provenance tracking, watermarking, and AI detection tools, are some of the current approaches being used to detect and counter these AI-generated risks. While these approaches can be beneficial for a variety of use cases, more research and investment are needed to improve their efficacy, security, and reliability. As NIST develops guidance and recommendations in this area, we encourage the Agency to review the Information Technology Industry's document on *Authenticating AI-Generated Content*[4] to inform its efforts. Noted above, NIST should consider impacts (e.g., bias) posed by content authentication techniques.

Additionally, multi-stakeholder approaches including technical standards and media education can help to reduce the risks of synthetic content. Public-private partnerships could join content creators, platforms, and companies to accelerate industry-led standardization efforts, such as the Coalition for Content Provenance and Authenticity[5] (C2PA), to help increase adoption of these specifications for provenance and authenticity of AI-generated content across stakeholders. For media education, the

[1] https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf.
[2] https://mneguidelines.oecd.org/.
[3] https://unesdoc.unesco.org/ark:/48223/pf0000381137.
[4] https://www.itic.org/policy/ITI_AIContentAuthorizationPolicy_122123.pdf.
[5] https://c2pa.org/.

development of new educational and training programs could equip content creators, journalists, and the public with skills and knowledge needed to understand and navigate the complex landscape of AI-generated information. Educating the public about potential risks and providing guidance on how to identify and report harms, can empower individuals to be more vigilant when engaging with generative AI applications.

Advance Responsible Global Technical Standards for AI Development

Potential Implications of Standards for Competition and International Trade

The development of international AI standards should be prioritized by NIST, the United States ("U.S.") government, and U.S. stakeholders for common standardization areas of interest across global markets. Aligned with the U.S. National Standards Strategy, focusing efforts on international standards increases U.S. influence and global leadership for safe, secure, and trustworthy AI – this is also competitively important as other countries increase investments to participate in and influence international AI standards.

For U.S. economic benefits, NIST and the United States government can role model for other countries the importance of prioritizing international AI standards development over development of unique national or regional standards which may fragment global markets, create market access barriers, or increase implementation burdens for the U.S. industry.   Furthermore, an emphasis on international standards development can best leverage limited expert resources by optimizing time and reducing costs.

NIST can further support these goals by:

1) To the extent possible, focusing NIST deliverables on the development of international AI standards, and contribute NIST deliverables to relevant international standards development organizations. With respect to the NIST AI RMF, further advances on the framework including on standards to support implementation of RMF functions would be best targeted as contributions to international standards such as ISO/IEC efforts. Given U.S. stakeholder interests

in both the NIST AI RMF and international standards efforts, prioritizing contributions to international work would also reduce resources and efforts needed to continue developing crosswalks between the NIST AI RMF (and other NIST deliverables) and relevant international standards.

2) Convening U.S. stakeholders (across public sector, private sector – multinational corporations, subject matter experts, academia, consortia etc.) on subject areas of focus for NIST pre-standardization work and development of tools and other resources to support creation and implementation of international standards. NIST's unique role can bring together a broader set of national stakeholders, which may not regularly engage in the formal international standards process – to develop U.S. consensus positions for inputs to U.S. international standards technical committees. This is in addition to increasing direct participation of NIST experts in U.S. international standards technical committees and coordinating the participation of experts in other United States government agencies.

3) Supporting U.S. stakeholders to adopt international standards by ensuring consistency between NIST-led deliverables with international standards. Aligning NIST deliverables with international standards and referencing relevant international standards to the extent possible would also be beneficial.

4) Prioritizing the development and use of international standards in bi-lateral and multi-lateral cooperation engagements with NIST counterparts in other countries.

5) Facilitating coordination of research investments for AI standardization pilot projects to accelerate the adoption of privacy-enhancing technologies and technical tools that facilitate bias and fairness detection and mitigation and continuous monitoring of AI system performance.

When setting generative AI standards, it will be important to define a set of risk assessment criteria for generative AI due to the variability of risks posed by different systems. This should accompany any set of specification and description language requirements based on the risk level and specific application. For specific security standards requirements, we recommend that NIST consider the areas listed below.

- Standards for use of AI to develop products: Guardrails and guidance around AI-generated code and designs, including manual (human) code review, IP scanning, static analysis, and security validation testing.
- Security for AI development: This includes threat modelling guidance for AI-specific threats such as model poisoning and prompt injection, in addition to supply chain vulnerabilities and access control.
- Defining recovery and resiliency capabilities, plans and policies: In the event of a compromise, the model can be safely shut down, or rolled back to an earlier "known good" state, depending on the nature of the incident.
- Defining security validation: This request for information refers to a number of controls, such as synthetic content labeling, data protection, model integrity monitoring, etc. Testing to ensure that those controls are in place and effective is especially important given the nascence of generative AI technology.
- Human training so that AI development teams are aware of the limitations, security, and ethical risks associated with generative AI. Understanding fundamentals of pattern recognition, what it can and cannot do, will assist them with mitigating and addressing problematic impacts (e.g., bias and confabulation).

Lastly, there are several organizations that have developed specific best practices, standards, and guidelines that should be considered when NIST undertakes advancing responsible standards for AI development. The Confidential Computing Consortium[6] (CCC) brings together hardware vendors, cloud providers, and software developers to accelerate the adoption of Trusted Execution Environment (TEE) technologies and standards. These can be applied to both privacy-enhancing-technologies, as well as better enable confidentiality and integrity of both AI algorithms and systems. MLCommons[7] is a non-profit consortium that aims to accelerate the benefits of machine learning and artificial intelligence, serving as a convener for the collaborative development of useful AI systems standards, including establishing and running

[6] https://confidentialcomputing.io/.

[7] https://mlcommons.org/.

successful benchmarks such as MLPerf, which characterizes the training and inference speed of AI hardware and software. The recently formed MLCommons AI Safety Working Group[8] is developing a platform and pool of tests to support AI safety benchmarks spanning a diverse set of use cases.

***

Intel welcomes the opportunity to discuss our feedback to this request for information. We look forward to continued engagement with NIST on this and other matters relating to the responsible development and deployment of AI technologies.

Sincerely,

Angel Preston
Policy Director, Artificial Intelligence
Intel Corporation

[8] https://mlcommons.org/working-groups/ai-safety/ai-safety/.