This document is scheduled to be published in the
Federal Register on 12/21/2023 and available online at
https://federalregister.gov/d/2023-28232, and on **https://govinfo.gov**

ode: 3510-13

DEPARTMENT OF COMMERCE

National Institute of Standards and Technology

[Docket Number: 231218-0309]

RIN: 0693-XC135

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)

AGENCY: National Institute of Standards and Technology (NIST), Commerce.

ACTION: Notice; Request for information.

SUMMARY: The National Institute of Standards and Technology (NIST) is seeking information to assist in carrying out several of its responsibilities under the Executive order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence issued on October 30, 2023. Among other things, 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.

DATES: Comments containing information in response to this notice must be received on or before February 2, 2024. Submissions received after that date may not be considered.

ADDRESSES:

Comments may be submitted by any of the following methods:

Electronic submission: Submit electronic public comments via the Federal e-Rulemaking Portal.

1. Go to *www.regulations.gov* and enter NIST–2023–0309 in the search field,
2. Click the "Comment Now!" icon, complete the required fields, and
3. Enter or attach your comments.

Electronic submissions may also be sent as an attachment to ai-inquiries@nist.gov and may be in any of the following unlocked formats: HTML; ASCII; Word; RTF; Unicode, or .pdf.

Written comments may also be submitted by mail to 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.

Response to this RFI is voluntary. Submissions must not exceed 25 pages (when printed) in 12-point or larger font, with a page number provided on each page. Please include your name, organization's name (if any), and cite “NIST AI Executive order” in all correspondence.

Comments containing references, studies, research, and other empirical data that are not widely published should include copies of the referenced materials. All comments and submissions, including attachments and other supporting materials, will become part of the public record and subject to public disclosure. Relevant comments will generally be available on the Federal eRulemaking Portal at *www.regulations.gov*. After the comment period closes, relevant comments will generally be available on *https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence*. NIST will not accept comments accompanied by a request that part or all of the material be treated confidentially because of its business proprietary nature or for any other reason. Therefore, do not submit confidential business information or otherwise sensitive, protected, or personal information, such as account numbers, Social Security numbers, or names of other individuals.

FOR FURTHER INFORMATION CONTACT: For questions about this RFI contact: ai-inquiries@nist.gov or Rachel Trello, National Institute of Standards and Technology, 100 Bureau Drive, Stop 8900, Gaithersburg, MD 20899, (202) 570-3978. Direct media inquiries to NIST's Office of Public Affairs at (301) 975-2762. Users of telecommunication devices for the deaf, or a text telephone, may call the Federal Relay Service toll free at 1-800-877-8339.

Accessible Format: NIST will make the RFI available in alternate formats, such as Braille or large print, upon request by persons with disabilities.

SUPPLEMENTARY INFORMATION: NIST is responsible for contributing to several deliverables assigned to the Secretary of Commerce. Among those is a report identifying existing standards, tools, methods, and practices, as well as the potential development of further science-backed and non-proprietary standards and techniques, related to synthetic content, including potentially harmful content, such as child sexual abuse material and non-consensual intimate imagery of actual adults. NIST will also assist the Secretary of Commerce to establish a plan for global engagement to promote and develop AI standards.

Respondents may provide information on one or more of the topics in this RFI and may elect not to address every topic.

NIST is seeking information to assist in carrying out several of its responsibilities under Sections 4.1, 4.5, and 11 of E.O. 14110. This RFI addresses the specific assignments cited below. Other assignments to NIST in E.O. 14110 related to cybersecurity and privacy, synthetic nucleic acid sequencing, and supporting agencies' implementation of minimum risk-management practices are being addressed separately. Information about NIST's assignments and plans under E.O. 14110, along with further opportunities for public input, may be found here: *https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence*.

In considering information for submission to NIST, respondents are encouraged to review recent guidance documents that NIST has developed with significant public input and feedback, including the NIST AI Risk Management Framework (*https://www.nist.gov/itl/ai-risk-management-framework*). Other NIST AI resources may be found on the NIST AI Resource Center (*https://airc.nist.gov/home*). In addition, respondents are encouraged to take into consideration the activities of the NIST Generative AI Public Working Group (*https://airc.nist.gov/generative_ai_wg*).

Information that is specific and actionable is of special interest, versus general statements about the challenges and needs. Copyright protections of materials, if any, should be clearly noted. Responses which include information generated by means of AI techniques should be identified clearly.

NIST is interested in receiving information pertinent to any or all of the assignments described below.

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

NIST is seeking information regarding topics related to generative AI risk management, AI evaluation, and red-teaming.

a. E.O. 14110 Sections 4.1(a)(i)(A) and (C) direct NIST to establish guidelines and best practices in order to promote consensus industry standards in the development and deployment of safe, secure, and trustworthy AI systems. Accordingly, NIST is seeking information regarding topics related to this assignment, including:

1) Developing a companion resource to the AI Risk Management Framework (AI RMF), NIST AI 100-1 (*https://www.nist.gov/itl/ai-risk-management-framework*), for generative AI. Following is a non-exhaustive list of possible topics that may be addressed in any comments relevant to AI RMF companion resource for generative AI:

- Risks and harms of generative AI, including challenges in mapping, measuring, and managing trustworthiness characteristics as defined in the AI RMF, as well as harms related to repression, interference with democratic processes and institutions, gender-based violence, and human rights abuses (see *https://www.whitehouse.gov/briefing-room/speeches-remarks/2023/11/01/remarks-by-vice-president-harris-on-the-future-of-artificial-intelligence-london-united-kingdom*);
- Current standards or industry norms or practices for implementing AI RMF core functions for generative AI (govern, map, measure, manage), or gaps in those standards, norms, or practices;
- Recommended changes for AI actors to make to their current governance practices to manage the risks of generative AI;

- The types of professions, skills, and disciplinary expertise organizations need to effectively govern generative AI, and what roles individuals bringing such knowledge could serve;
- 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);
- Current techniques and implementations, including their feasibility, validity, fitness for purpose, and scalability, for:
    - Model validation and verification, including AI red-teaming;
    - Human rights impact assessments, ethical assessments, and other tools for identifying impacts of generative AI systems and mitigations for negative impacts;
    - Content authentication, provenance tracking, and synthetic content labeling and detection, as described in Section 2a below; and
    - Measurable and repeatable mechanisms to assess or verify the effectiveness of such techniques and implementations.
- Forms of transparency and documentation (*e.g.*, model cards, data cards, system cards, benchmarking results, impact assessments, or other kinds of transparency reports) that are more or less helpful for various risk management purposes (*e.g.,* assessment, evaluation, monitoring, and provision of redress and contestation mechanisms) and for various AI actors (developers, deployers, end users, etc.) in the context of generative AI models, and best practices to ensure such information is shared as needed along the generative AI lifecycle and supply chain);
- Economic and security implications of watermarking, provenance tracking, and other content authentication tools;

- Efficacy, validity, and long-term stability of watermarking techniques and content authentication tools for provenance of materials, including in derivative work;
- Criteria for defining an error, incident, or negative impact;
- Governance policies and technical requirements for tracing and disclosing errors, incidents, or negative impacts;
- The need for greater controls when data are aggregated; and
- The possibility for checks and controls before applications are presented forward for public consumption.

2) Creating guidance and benchmarks for evaluating and auditing AI capabilities, with a focus on capabilities and limitations through which AI could be used to cause harm. Following is a non-exhaustive list of possible topics that may be addressed in any comments relevant to AI evaluations:
    - Definition, types, and design of test environments, scenarios, and tools for evaluating the capabilities, limitations, and safety of AI technologies;
    - Availability of, gap analysis of, and proposals for metrics, benchmarks, protocols, and methods for measuring AI systems' functionality, capabilities, limitations, safety, security, privacy, effectiveness, suitability, equity, and trustworthiness. This includes rigorous measurement approaches for risks and impacts such as:
        - Negative effects of system interaction and tool use, including from the capacity to control physical systems or from reliability issues with such capacity or other limitations;
        - Exacerbating chemical, biological, radiological, and nuclear (CBRN) risks;

    - Enhancing or otherwise affecting malign cyber actors' capabilities, such as by aiding vulnerability discovery, exploitation, or operational use;
    - Introduction of biases into data, models, and AI lifecycle practices;
    - Risks arising from AI value chains in which one developer further refines a model developed by another, especially in safety- and rights-affecting systems;
    - Impacts to human and AI teaming performance;
    - Impacts on equity, including such issues as accessibility and human rights; and
    - Impacts to individuals and society; including both positive and negative impacts on safety and rights.
- Generalizability of standards and methods of evaluating AI over time, across sectors, and across use cases;
- Applicability of testing paradigms for AI system functionality, effectiveness, safety, and trustworthiness including security, and transparency, including paradigms for comparing AI systems against each other, baseline system performance, and existing practice, such as:
    - Model benchmarking and testing; and
    - Structured mechanisms for gathering human feedback, including randomized controlled human-subject trials; field testing, A/B testing, AI red-teaming.

b. E.O. 14110 Section 4.1(a)(ii) directs NIST to establish guidelines (except for AI used as a component of a national security system), including appropriate procedures and processes, to enable developers of AI, especially of dual-use foundation models, to conduct AI red-teaming tests to enable deployment of safe, secure, and trustworthy systems. The following is a non-exhaustive list of possible topics that may be addressed in any comments relevant to red-teaming:

- Use cases where AI red-teaming would be most beneficial for AI risk assessment and management;
- Capabilities, limitations, risks, and harms that AI red-teaming can help identify considering possible dependencies such as degree of access to AI systems and relevant data;
- Current red-teaming best practices for AI safety, including identifying threat models and associated limitations or harmful or dangerous capabilities;
- Internal and external review across the different stages of AI life cycle that are needed for effective AI red-teaming;
- Limitations of red-teaming and additional practices that can fill identified gaps;
- Sequence of actions for AI red-teaming exercises and accompanying necessary documentation practices;
- Information sharing best practices for generative AI, including for how to share with external parties for the purpose of AI red-teaming while protecting intellectual property, privacy, and security of an AI system;
- How AI red-teaming can complement other risk identification and evaluation techniques for AI models;
- How to design AI red-teaming exercises for different types of model risks, including specific security risks (e.g., CBRN risks, etc.) and risks to individuals and society (e.g., discriminatory output, hallucinations, etc.);
- Guidance on the optimal composition of AI red teams including different backgrounds and varying levels of skill and expertise;
- Economic feasibility of conducting AI red-teaming exercises for small and large organizations; and

- The appropriate unit of analysis for red teaming (models, systems, deployments, etc.)

2. **Reducing the Risk of Synthetic Content**

NIST is seeking information regarding topics related to synthetic content creation, detection, labeling, and auditing.

a. E.O. 14110 Section 4.5(a) 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. NIST is seeking information regarding the following topics related to reducing the risk of synthetic content in both closed and open source models that should be included in the Secretary's report, recognizing that the most promising approaches will require multistakeholder input, including scientists and researchers, civil society, and the private sector. Existing tools and the potential development of future tools, measurement methods, best practices, active standards work, exploratory approaches, challenges and gaps are of interest for the following non-exhaustive list of possible topics and use cases of particular interest.

- Authenticating content and tracking its provenance;
- Techniques for labeling synthetic content, such as using watermarking;
- Detecting synthetic content;
- Resilience of techniques for labeling synthetic content to content manipulation;
- Economic feasibility of adopting such techniques for small and large organizations;
- Preventing generative AI from producing child sexual abuse material or producing non-consensual intimate imagery of real individuals (to include intimate digital depictions of the body or body parts of an identifiable individual);

- Ability for malign actors to circumvent such techniques;
- Different risk profiles and considerations for synthetic content for models with widely available model weights;
- Approaches that are applicable across different parts of the AI development and deployment lifecycle (including training data curation and filtering, training processes, fine-tuning incorporating both automated means and human feedback, and model release), at different levels of the AI system (including the model, API, and application level), and in different modes of model deployment (online services, within applications, open-source models, etc.);
- Testing software used for the above purposes; and
- Auditing and maintaining tools for analyzing synthetic content labeling and authentication.

3. Advance responsible global technical standards for AI development

NIST is seeking information regarding topics related to the development and implementation of AI-related consensus standards, cooperation and coordination, and information sharing that should be considered in the design of standards.

a. E.O. 14110 Section 11(b) directs the Secretary of Commerce, within 270 days and in coordination with the Secretary of State and the heads of other relevant agencies, 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 (*https://www.nist.gov/itl/ai-risk-management-framework*) and the U.S. Government National Standards Strategy for Critical and Emerging Technology (*https://www.whitehouse.gov/wp-content/uploads/2023/05/US-Gov-National-Standards-Strategy-2023.pdf*). The following is a non-exhaustive list of possible topics that may be addressed:

- AI nomenclature and terminology;
- Best practices regarding data capture, processing, protection, quality, privacy, transparency, confidentiality, handling, and analysis, as well as inclusivity, fairness, accountability, and representativeness (including non-discrimination, representation of lower resourced languages, and the need for data to reflect freedom of expression) in the collection and use of data;
- Examples and typologies of AI systems for which standards would be particularly impactful (e.g., because they are especially likely to be deployed or distributed across jurisdictional lines, or to need special governance practices);
- Best practices for AI model training;
- Guidelines and standards for trustworthiness, verification, and assurance of AI systems;
- AI risk management and governance, including managing potential risk and harms to people, organizations, and ecosystems;
- Human-computer interface design for AI systems;
- Application specific standards (e.g., for computer vision, facial recognition technology);
- Ways to improve the inclusivity of stakeholder representation in the standards development process;
- Suggestions for AI-related standards development activities, including existing processes to contribute to and gaps in the current standards landscape that could be addressed, and including with reference to particular impacts of AI;
- Strategies for driving adoption and implementation of AI-related international standards;
- Potential mechanisms, venues, and partners for promoting international collaboration, coordination, and information sharing on standards development;

- Potential implications of standards for competition and international trade; and
- Ways of tracking and assessing whether international engagements under the plan are having the desired impacts.

Across all these topics, NIST is seeking information about costs and ease of implementation for tools, systems, practices, and the extent to which they will benefit the public if they can be efficiently adopted and utilized.

Authority: Executive Order 14110 of Oct. 30, 2023; 15 U.S.C. 272.

Alicia Chambers,

NIST Executive Secretariat.

[FR Doc. 2023-28232 Filed: 12/19/2023 4:15 pm; Publication Date: 12/21/2023]