3510-13

DEPARTMENT OF COMMERCE

National Institute of Standards and Technology

[Docket Number: [210726-0151]]

Artificial Intelligence Risk Management Framework

AGENCY: National Institute of Standards and Technology, Department of Commerce.

ACTION: Request for Information.

SUMMARY: The National Institute of Standards and Technology (NIST) is developing a framework that can be used to improve the management of risks to individuals, organizations, and society associated with artificial intelligence (AI). The NIST Artificial Intelligence Risk Management Framework (AI RMF or Framework) is intended for voluntary use and to improve the ability to incorporate trustworthiness considerations into the design, development, and use, and evaluation of AI products, services, and systems. This notice requests information to help inform, refine, and guide the development of the AI RMF. The Framework will be developed through a consensus-driven, open, and collaborative process that will include public workshops and other opportunities for stakeholders to provide input.

DATES: Comments in response to this notice must be received by 5:00 p.m. Eastern time on August 19, 2021. Written comments in response to the RFI should be submitted according to the instructions in the ADDRESSES and SUPPLEMENTARY INFORMATION sections below. 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-2021-0004 in the search field,
    2. Click the “Comment Now!” icon, complete the required fields, and

3. Enter or attach your comments.

- *Email*: Comments in electronic form may also be sent to AIframework@nist.gov in any of the following formats: HTML; ASCII; Word; RTF; or PDF.

Please submit comments only and include your name, organization's name (if any), and cite "AI Risk Management Framework" in all correspondence.

FOR FURTHER INFORMATION CONTACT: For questions about this RFI contact: Mark Przybocki (mark.przybocki@nist.gov), U.S. National Institute of Standards and Technology, MS 20899, 100 Bureau Drive, Gaithersburg, MD 20899, telephone (301) 975-3347, email AIframework@nist.gov.

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: On request to the contact person listed above, NIST will make the RFI available in alternate formats, such as Braille or large print, upon request by persons with disabilities.

SUPPLEMENTARY INFORMATION:

Genesis for Development of the AI Risk Management Framework

Artificial intelligence (AI) is rapidly transforming our world.

Surges in AI capabilities have led to a wide range of innovations. These new AI-enabled systems are benefitting many parts of society and economy from commerce and healthcare to transportation and cybersecurity. At the same time, new AI-based technologies, products, and services bring technical and societal challenges and risks, including ensuring that AI comports with ethical values. While there is no objective standard for ethical values, as they are grounded in the norms and legal expectations of specific societies or cultures, it is widely agreed that AI must be designed, developed, used, and evaluated in a trustworthy and responsible manner to foster public confidence

and trust. Trust is established by ensuring that AI systems are cognizant of and are built to align with core values in society, and in ways which minimize harms to individuals, groups, communities, and societies at large.

Defining trustworthiness in meaningful, actionable, and testable ways remains a work in progress. Inside and outside the United States there are diverse views about what that entails, including who is responsible for instilling trustworthiness during the stages of design, development,use, and evaluation. There also are different ideas about how to assure conformity with principles and characteristics of AI trustworthiness.

NIST is among the institutions addressing these issues. NIST aims to cultivate the public's trust in the design, development, use, and evaluation of AI technologies and systems in ways that enhance economic security, and improve quality of life. NIST focuses on improving measurement science, standards, technology, and related tools, including evaluation and data. NIST is developing forward-thinking approaches that support innovation and confidence in AI systems. The agency's work on an AI RMF is consistent with recommendations by the National Security Commission on Artificial Intelligence[1] and the Plan for Federal Engagement in Developing AI Technical Standards and Related Tools.[2]

Congress has directed NIST to collaborate with the private and public sectors to develop a voluntary AI RMF.[3] The Framework is intended to help designers, developers, users and evaluators of AI systems better manage risks across the AI lifecycle. For purposes of this RFI, "managing" means: identifying, assessing, responding to, and

---

[1] National Security Commission on Artificial Intelligence, Final Report, https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.

[2] Plan for Federal Engagement in Developing AI Technical Standards and Related Tools, https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf

[3] H. Rept. 116-455 - COMMERCE, JUSTICE, SCIENCE, AND RELATED AGENCIES APPROPRIATIONS BILL, 2021, CRPT-116hrpt455.pdf (congress.gov), and Section 5301 of the National Artificial Intelligence Initiative Act of 2020 (Pub. L. No. 116-283), https://www.congress.gov/116/bills/hr6395/BILLS-116hr6395enr.pdf.

communicating AI risks. “Responding” to AI risks means: avoiding, mitigating, sharing, transferring, or accepting risk. “Communicating” AI risk means: disclosing and negotiating risk and sharing with connected systems and actors in the domain of design, deployment and use. “Design, development, use, and evaluation” of AI systems includes procurement, monitoring, or sustainment of AI components and systems.

The Framework aims to foster the development of innovative approaches to address characteristics of trustworthiness including accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of unintended and/or harmful bias, as well as of harmful uses. The Framework should consider and encompass principles such as transparency, fairness, and accountability during design, deployment, use, and evaluation of AI technologies and systems. With broad and complex uses of AI, the Framework should consider risks from unintentional, unanticipated, or harmful outcomes that arise from intended uses, secondary uses, and misuses of the AI. These characteristics and principles are generally considered as contributing to the trustworthiness of AI technologies and systems, products, and services. NIST is interested in whether stakeholders define or use other characteristics and principles.

Among other purposes, the AI RMF is intended to be a tool that would complement and assist with broader aspects of enterprise risk management which could affect individuals, groups, organizations, or society.

AI RMF Development and Attributes

NIST is soliciting input from all interested stakeholders, seeking to understand how individuals, groups and organizations involved with designing, developing, using, or evaluating AI systems might be better able to address the full scope of AI risk and how a framework for managing AI risks might be constructed. Stakeholders include but are not

limited to industry, civil society groups, academic institutions, federal agencies, state, local, territorial, tribal, and foreign governments, standards developing organizations and researchers.

NIST intends the Framework to provide a prioritized, flexible, risk-based, outcome-focused, and cost-effective approach that is useful to the community of AI designers, developers, users, evaluators, and other decision makers and is likely to be widely adopted. The Framework's development process will involve several iterations to encourage robust and continuing engagement and collaboration with interested stakeholders. This will include open, public workshops, along with other forms of outreach and feedback. This RFI is an important part of that process.

NIST believes that the AI RMF should have the following attributes:

1. Be consensus-driven and developed and regularly updated through an open, transparent process. All stakeholders should have the opportunity to contribute to the Framework's development. NIST has a long track record of successfully and collaboratively working with a range of stakeholders to develop standards and guidelines. NIST will model its approach on the open, transparent, and collaborative approaches used to develop the Framework for Improving Critical Infrastructure Cybersecurity ("Cybersecurity Framework")[4] as well as the Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management ("Privacy Framework").[5]

2. Provide common definitions. The Framework should provide definitions and characterizations for aspects of AI risk and trustworthiness that are common and relevant

[4] Framework for Improving Critical Infrastructure Cybersecurity ("Cybersecurity Framework"), https://www.nist.gov/cyberframework.

[5] Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management ("Privacy Framework"), https://www.nist.gov/privacy-framework/privacy-framework.

across all sectors. The Framework should establish common AI risk taxonomy, terminology, and agreed- upon definitions, including that of trust and trustworthiness.

3. Use plain language that is understandable by a broad audience, including senior executives and those who are not AI professionals, while still of sufficient technical depth to be useful to practitioners across many domains.

4. Be adaptable to many different organizations, AI technologies, lifecycle phases, sectors, and uses. The Framework should be scalable to organizations of all sizes, public or private, in any sector, and operating within or across domestic borders. It should be platform- and technology- agnostic and customizable. It should meet the needs of AI designers, developers, users, and evaluators alike.

5. Be risk-based, outcome-focused, voluntary, and non-prescriptive. The Framework should focus on the value of trustworthiness and related needs, capabilities, and outcomes. It should provide a catalog of outcomes and approaches to be used voluntarily, rather than a set of one-size-fits-all requirements, in order to: foster innovation in design, development, use and evaluation of trustworthy and responsible AI systems; inform education and workforce development; and promote research on and adoption of effective solutions. The Framework should assist those designing, developing, using, and evaluating AI to better manage AI risks for their intended use cases or scenarios.

6. Be readily usable as part of any enterprise's broader risk management strategy and processes.

7. Be consistent, to the extent possible, with other approaches to managing AI risk. The Framework should, when possible, take advantage of and provide greater awareness of existing standards, guidelines, best practices, methodologies, and tools for managing AI risks whether presented as frameworks or in other formats. It should be law- and

regulation-agnostic to support organizations' ability to operate under applicable domestic and international legal or regulatory regimes.

8. Be a living document. The Framework should be capable of being readily updated as technology, understanding, and approaches to AI trustworthiness and uses of AI change and as stakeholders learn from implementing AI risk management. NIST expects there may be aspects of AI trustworthiness that are not sufficiently developed for inclusion in the initial Framework.

As noted below, NIST solicits comments on these and potentially other desired attributes of an AI RMF, as well as on high-priority gaps in organizations' ability to manage AI risks.

Goals of This Request for Information (RFI)

This RFI invites stakeholders to submit ideas, based on their experience as well as their research, to assist in prioritizing elements and development of the AI RMF. Stakeholders include but are not limited to industry, civil society groups, academic institutions, federal agencies, state, local, territorial, tribal, and foreign governments, standards developing organizations and researchers. The Framework is intended to address AI risk management related to individuals, groups or organizations involved in the design, development, use, and evaluation of AI systems.

The goals of the Framework development process, generally, and this RFI, specifically, are to:

1. identify and better understand common challenges in the design, development, use, and evaluation of AI systems that might be addressed through a voluntary Framework;

2. gain a greater awareness about the extent to which organizations are identifying, assessing, prioritizing, responding to, and communicating AI risk or have incorporated AI risk management standards, guidelines, and best practices, into their policies and practices; and

3. specify high-priority gaps for which guidelines, best practices, and new or revised standards are needed and could be addressed by the AI RMF – or which would require further understanding, research, and development.

Details About Responses to This Request for Information

When addressing the topics below, respondents may describe the practices of their organization or organizations with which they are familiar. They also may provide information about the type, size, and location of those organization(s) if they desire. Providing such information is optional and will not affect NIST's full consideration of the comment. Respondents are encouraged to provide generalized information based on research and potential practices as well as on current approaches and activities. Comments containing references, studies, research, and other empirical data that are not widely published (e.g., available on the internet) should include copies of the referenced materials. All submissions, including attachments and other supporting materials, will become part of the public record and subject to public disclosure. NIST reserves the right to publish relevant comments publicly, unedited and in their entirety. All relevant comments received by the deadline will be made publicly available at https://www.nist.gov/itl/ai-risk-management-framework and at regulations.gov. Respondents are strongly encouraged to use the template available at: https://www.nist.gov/itl/ai-risk-management-framework.

Personally identifiable information (PII), such as street addresses, phone numbers, account numbers or Social Security numbers, or names of other individuals, should not be included. NIST asks commenters to avoid including PII as NIST has no plans to redact PII from comments. Do not submit confidential business information, or otherwise sensitive or protected information. Comments that contain profanity, vulgarity, threats, or other inappropriate language or content will not be considered. NIST requests that commenters, to the best of their ability, only submit attachments that are accessible to people who rely upon assistive technology. A good resource for document accessibility can be found at: section508.gov/create/documents.

Specific Requests for Information

The following statements are not intended to limit the topics that may be addressed. Responses may include any topic believed to have implications for the development of an AI RMF, regardless of whether the topic is included in this document. All relevant responses that comply with the requirements listed in the DATES and ADDRESSES sections of this RFI and set forth below will be considered.

NIST is requesting information related to the following topics:

1. The greatest challenges in improving how AI actors manage AI-related risks – where "manage" means identify, assess, prioritize, respond to, or communicate those risks;
2. How organizations currently define and manage characteristics of AI trustworthiness and whether there are important characteristics which should be considered in the Framework besides: accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of harmful bias, or harmful outcomes from misuse of the AI;

3. How organizations currently define and manage principles of AI trustworthiness and whether there are important principles which should be considered in the Framework besides: transparency, fairness, and accountability;
4. The extent to which AI risks are incorporated into different organizations' overarching enterprise risk management – including, but not limited to, the management of risks related to cybersecurity, privacy, and safety;
5. Standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles to identify, assess, prioritize, mitigate, or communicate AI risk and whether any currently meet the minimum attributes described above;
6. How current regulatory or regulatory reporting requirements (e.g., local, state, national, international) relate to the use of AI standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles;
7. AI risk management standards, frameworks, models, methodologies, tools, guidelines and best practices, principles, and practices which NIST should consider to ensure that the AI RMF aligns with and supports other efforts;
8. How organizations take into account benefits and issues related to inclusiveness in AI design, development, use and evaluation – and how AI design and development may be carried out in a way that reduces or manages the risk of potential negative impact on individuals, groups, and society.
9. The appropriateness of the attributes NIST has developed for the AI Risk Management Framework. (See above, "AI RMF Development and Attributes");

10. Effective ways to structure the Framework to achieve the desired goals, including, but not limited to, integrating AI risk management processes with organizational processes for developing products and services for better outcomes in terms of trustworthiness and management of AI risks. Respondents are asked to identify any current models which would be effective. These could include – but are not limited to – the NIST Cybersecurity Framework or Privacy Framework, which focus on outcomes, functions, categories and subcategories and also offer options for developing profiles reflecting current and desired approaches as well as tiers to describe degree of framework implementation; and

11. How the Framework could be developed to advance the recruitment, hiring, development, and retention of a knowledgeable and skilled workforce necessary to perform AI-related functions within organizations.

12. The extent to which the Framework should include governance issues, including but not limited to make up of design and development teams, monitoring and evaluation, and grievance and redress.

Authority: 15 U.S.C. 272(b), (c), & (e); 15 U.S.C. 278g-3.

Alicia Chambers,
NIST Executive Secretariat.
[FR Doc. 2021-16176 Filed: 7/28/2021 8:45 am; Publication Date: 7/29/2021]