KAISER PERMANENTE®

February 2, 2024

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

Submitted electronically to: www.regulations.gov

RE: Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence [NIST–2023–0009–0001]

Dear Ms. Trello:

Kaiser Permanente offers the following comments on the above-captioned RFI. Kaiser Permanente[1] is the largest private integrated health care delivery system in the United States, with more than 12.6 million members in eight states and the District of Columbia. Our mission is to provide high-quality, affordable health care services and to improve the health of our members and the communities we serve.

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

A. NIST seeks input related to generative AI risk management, AI evaluation, and red-teaming. Executive Order (E.O.) 14110[2] directs NIST to establish guidelines and best practices to promote consensus industry standards in the development and deployment of safe, secure, and trustworthy AI systems, including the following topics.

Developing a companion resource to the AI Risk Management Framework (AI RMF), NIST AI 100-1, for Generative AI

NIST proposes to include a broad array of topics including but not limited to:

- Associated risks and harms.
- Current implementation standards, industry norms, or practices
- Changes to current governance practices to manage risks
- Expertise needed for effective governance

[1] Kaiser Permanente comprises Kaiser Foundation Health Plan, Inc., one of the nation's largest not-for-profit health plans, and its health plan subsidiaries outside California and Hawaii; the not-for-profit Kaiser Foundation Hospitals, which operates 40 hospitals and more than 600 other clinical facilities; and the Permanente Medical Groups, self-governed physician group practices that exclusively contract with Kaiser Foundation Health Plan and its health plan subsidiaries to meet the health needs of Kaiser Permanente's members.

[2] Executive Office of the President. *Safe, Secure and Trustworthy Development and Use of Artificial Intelligence*. Sections 4.1(a)(i)(A) and (C), October 30, 2023. 88 Fed. Reg. 75191. *Available at* https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

- Roles for managing risks and harms
- Current techniques and implementations
- Forms of transparency and documentation
- Provenance/authentication
- Error definition/disclosure/governance
- Data aggregation
- Checks/controls prior to public consumption

Given the current stage of AI adoption and how rapidly AI is expected to evolve and mature, the AI RMF is an important activity NIST should undertake in the short term. We recommend NIST first finalize and issue the AI RMF Version 2.0, as Version 1.0 is too early a draft to provide significant guidance. In addition to the topics listed above, we also recommend NIST consider including:

- Industry-specific guidance, e.g., for the health care sector and other critical infrastructure sectors
- Information privacy and security policies' impact on AI development and use (large databases, individual-level data, privacy protections, consumer informed consent, access controls, etc.)

More generally, and to enhance value and consistency, the NIST AI Risk Management Framework and NIST's coordination efforts with other offices and agencies should align across other requirements and supporting frameworks (e.g., the Cybersecurity Maturity Model Certification (CMMC), the NIST Enterprise Risk Management Framework (ERM), the NIST Cybersecurity Framework, and the NIST Privacy Framework). It will be critical to establish a common federal technical approach to key concepts and definitions such as data quality, bias, explainability and auditability.

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

NIST proposes possible topics for comment, including but not limited to:

- Definition, types, and design of test environments, scenarios, and tools for evaluating capabilities, limitations, and safety
- 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
- 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

While these topics are important ones on which to solicit feedback, we wish to note that the type of AI and use cases, as well as outcomes, will have an impact on the applicability of these topics and the ability to comply with related requirements. As AI depends on large volumes of high-quality data, model accuracy and outcomes will directly reflect the training data. Therefore, data governance and data integrity controls will be critical to AI outcomes, assessment, analysis, outcomes, and controls.

There will also be a direct impact to data privacy and those regulations and concerns. The U.S. does not currently have a pre-emptive national data privacy law (unlike E.U. countries, which rely on the General Data Protection Regulation (GDPR)). The Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule applies to some data in some AI systems of specified entities. Numerous states have enacted data privacy laws (and other AI-related regulations) that will interact or conflict with any federal-level AI regulation. This will require Kaiser Permanente and other organizations to map internal practices, controls, and policies to both state and federal requirements. Additionally, as noted above, use cases per industry will be key to regulated entities' ability to align and map these risks to outcomes and controls. We reiterate our comment about the importance and value of creating sector-specific use case guidance.

Kaiser Permanente strongly supports the proposal to establish metrics to rigorously measure AI systems' functionality, limitations, safety, security, privacy, effectiveness, suitability, equity, and trustworthiness. We specifically call attention to the value of measuring both the positive and negative impacts and effects of AI use, particularly with vulnerable populations.

We also recommend that Generative AI vendors provide a platform for an end user organization to benchmark their AI products' quality and safety, both at the outset and over time. It is not feasible to depend on end users to develop this type of functionality independently.

Additionally, shutdown/downtime protocols are needed when AI either has outages (akin to procedures that occur when an EHR is down) or starts exhibiting anomalous behavior. It is not sufficient for regulated entities to have only a kill-switch; there must also be procedures to handle the dependencies between AI systems and other systems and procedures.

B. E.O. 14110 Section 4.1(a)(ii) directs NIST to establish guidelines, 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. NIST seeks comments on topics relevant to red-teaming, including but not limited to:

- Use cases where AI red-teaming would be most beneficial.
- Capabilities, limitations, risks, and harms that AI red-teaming can help identify.
- Current red-teaming best practices for AI safety.
- Internal and external review across the different stages of AI life cycles needed for effective AI red-teaming.
- Limitations of red-teaming and additional practices that can fill identified gaps.
- Action sequences of AI red-teaming exercises and documentation practices.
- Best practices of information sharing for generative AI.
- How AI red-teaming can complement other risk identification and evaluation techniques.
- How to design AI red-teaming exercises for different types of model risks.
- Guidance on the optimal composition of AI red teams.
- Economic feasibility of AI red-teaming exercises for small and large organizations.
- The appropriate unit of analysis for red teaming (models, systems, deployments, etc.).

Kaiser Permanente agrees that NIST guidelines for red-teaming AI applications will be very helpful. It would also be valuable to provide sector-specific guidelines to enhance the ability of various sectors to use red-teaming effectively. For instance, in health care, the use of red-teaming for identifying risks,

threats, and negative clinical and other health care effects on individuals and populations, although important, can be more challenging to perform. In a health care setting, simulation of high-stakes emergencies improves team functioning and patient outcomes. AI-human simulation training will be needed to ensure that handoffs and gaps are well understood.

Therefore, in addition to the list of topics identified in this section of the RFI, NIST should also consider explicitly highlighting cybersecurity risks. We recommend adding "Red-teaming approaches and best practices to identify AI cybersecurity vulnerabilities and threats in the various models being used" as a topic in this section. Additional guidance could help organizations optimize red-teaming under different risk scenarios.

Part 2: Reducing the Risk of Synthetic Content

A. NIST is seeking information regarding topics related to synthetic content creation, detection, labeling, and auditing. E.O. 14110 Section 4.5(a) directs the Secretary to submit a report 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 from stakeholders regarding the following topics, including but not limited to:

- Authenticating content and tracking its provenance
- Techniques for labeling synthetic content, e.g., 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
- 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, at different levels of the AI system, and in different modes of model deployment
- Testing software used for the above purposes
- Auditing and maintaining tools for analyzing synthetic content labeling and authentication

Kaiser Permanente recommends establishing a minimum set of defined characteristics that synthetic content is intended to mimic or represent to help determine whether results accurately represent the referenced content. For example, such references for given populations could include the distribution of various demographic and socio-economic factors (such as race, ethnicity, age, geographic location, economic factors, education level, etc.)

It will also be important to establish national (or international) standards for an unremovable, unmodifiable watermark or label to accompany synthetic data.

Scope is an important consideration at the technical level: NIST should consider the model development phase within scope as well as applications and IT infrastructure on which AI runs. Scope could expand to encompass AI data centers and leverage systems that support AI critical infrastructure.

Part 3: Advance Responsible Global Technical Standards for AI Development

A. NIST seeks comments about the development and implementation of AI-related consensus standards, cooperation and coordination, and information sharing that should be considered in the design of standards. E.O. 14110 Section 11(b) directs the Secretary 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[3] and the U.S. Government National Standards Strategy for Critical and Emerging Technology.[4] The following is a non-exhaustive list of topics to address.

- 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 in the collection and use of data
- Examples and typologies of AI systems for which standards would be particularly impactful
- Best practices for AI model training
- Guidelines and standards for trustworthiness, verification, and assurance of AI systems
- AI risk management and governance
- Human-computer interface design for AI systems
- Application specific standards
- Ways to improve the inclusivity of stakeholder representation in the standards development process
- Suggestions for AI-related standards development activities
- 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
- Ways of tracking and assessing the impact of international engagements under the plan

We recommend NIST participate in and lead U.S. stakeholder efforts in national and international voluntary consensus standards development for the above areas through designated or accredited groups (e.g., the International Organization for Standardization/International Electrotechnical Commission

[3] AI Risk Management Framework (AI RMF 1.0), *available at* https://www.nist.gov/itl/ai-risk-management-framework (January 26,2023)
[4] United States Government National Standards Strategy for Critical and Emerging Technology, *available at* https://www.whitehouse.gov/wp-content/uploads/2023/05/US-Gov-National-Standards-Strategy-2023.pdf (May 2023)

(ISO/IEC) Joint Technical Committee 1 Subcommittee 42 (AI) and the InterNational Committee for Information Technology Standards (INCITS) AI work groups, or the Institute of Electrical and Electronics Engineers (IEEE) Standards Association Artificial Intelligence Standards Committee). Strict alignment of all aspects of AI standards development with the World Trade Organization Technical Barriers to Trade (WTO TBT) Agreement processes for collaboration and cooperation will be the best way to meet aims stated by NIST and in the E.O. and avoid adverse effects of conflicting standards.

International harmonization will be imperative for effective and ethical AI development and use across countries and within the global economy. Most countries are currently considering AI's impact and debating how to promote and regulate its use. Different national rules will inevitably conflict to some degree. While U.S.-based organizations like Kaiser Permanente may not be immediately affected, the downstream effects of global AI will eventually become pervasive, thus assessment and planning should start now.

For example, the G7 has launched the "Hiroshima AI process," the Organisation for Economic Co-operation and Development (OECD) has developed AI principles, and the United Nations has proposed a new UN AI Advisory board. Additionally, international technical standards (e.g., ISO/IEC/IEEE) also promise to propose their own approaches. NIST should develop a broad understanding of these various initiatives in deciding how it can best achieve assessment and harmonization.

Kaiser Permanente agrees that international cooperation in the establishment of standards and guidance for the safe, secure, effective and transparent use of AI is vitally important. We recommend adding these five areas to the list included for global standards cooperation and collaboration:

- Privacy – Guidelines and standards for the use of AI consistent with information privacy policy
- Cybersecurity – Guidelines and standards for protecting AI models from cybersecurity threats and vulnerabilities (Cybersecurity by Design)
- Equity – Guidelines and standards for ensuring AI tools do not discriminate or result in adverse effects on vulnerable populations
- Education – Best practices and effective approaches to implement communication and education campaigns to various audiences regarding the use of AI, its benefits and risks
- Workforce Development – Effective strategies to advance workforce development on AI

Thank you for considering these comments. Please contact Jamie Ferguson (jamie.ferguson@kp.org) or Lori Potter (lori.potter@kp.org) if we may provide additional information or answer any questions.

Sincerely,

Jamie Ferguson
Vice President
Health IT Strategy & Policy