CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY | GEORGETOWN UNIVERSITY

February 2, 2024

Agency name: National Institute of Standards and Technology
Federal Register Document Citation: 88 FR 88368
Organization: The Center for Security and Emerging Technology (CSET)
Respondent type: Organization>Academic institution / Think tank
Primary POC: Mina Narayanan (mjn82@georgetown.edu)

The Center for Security and Emerging Technology (CSET) at Georgetown University offers the following comments in response to the 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). A policy research organization within Georgetown University's Walsh School of Foreign Service, CSET produces data-driven research at the intersection of security and technology, providing nonpartisan analysis to the policy community. We appreciate the opportunity to offer these comments.

We have organized our feedback according to six topics featured in the RFI:

1. red-teaming;
2. criteria for defining an error, incident, or negative impact; and governance policies;
3. technical requirements for managing errors, incidents, or negative impacts;
4. AI risk management and governance;
5. strategies for driving adoption and implementation of AI standards; and,
6. potential mechanisms, venues, and partners for promoting standards development.

Where feedback is informed by or directly pulled from published research, we include hyperlinks to the relevant CSET publications.

Red-teaming

Relevance: NIST RFI Section 1 on *Developing Guidelines, Standards, and Best Practices for AI Safety and Security*; E.O. 14110 Section 4.1(a)(ii)
Referenced Publications: What Does AI Red-Teaming Actually Mean?

- AI red-teaming practices should not be limited to prompt-based testing. Red-teaming as performed in the cybersecurity context involves multiple types of adversarial testing at both the application and the system level, and guidelines for AI red-teaming should incorporate a more holistic approach that incorporates both safety and security concerns.
    - In addition to specific types of model risks such as discriminatory output, hallucinations, and/or hazardous information risks, cybersecurity-focused red-teaming should be conducted at multiple levels: model, system, user-machine interaction, and operational deployment/application. Adversarial testing at multiple levels can help inform assessments of how risks might arise from interactions between systems (e.g., how a model's behavior might be impacted if a user can provide backdoor inputs that bypass a safety or moderation control, or how a model with access to sensitive data might leak that information in certain contexts).
    - There is a general lack of understanding of how unique vulnerabilities or risks may arise when a foundation model is embedded into a larger system and given various privileges. Red-teaming practices should prioritize identifying system and deployment vulnerabilities, especially for potential U.S. government applications, as these vulnerabilities are currently less well documented and understood.
- Red-teaming can be divided into three rough categories depending on who is performing the testing. For foundation models, frontier models, and safety-critical AI systems, all three categories of red-teaming should be executed. Those categories are:
    - Red-teaming designed, directed, and executed by employees of the organization developing AI.
    - Red-teaming designed and directed by employees of the organization developing AI (with input from outside experts that may influence test

design and execution), and execution of the red-teaming by outside experts.
    - Red-teaming that is fully designed, directed, executed, and overseen by an independent outside organization. The developer provides the access needed for red-teaming but is otherwise not involved.
- Construction of cyber ranges or testbeds specifically for foundation model red-teaming (or adversarial testing) could be a valuable investment and allow for longer-term research as system capabilities improve.
- Designing AI red-teaming exercises tailored to specific risks will require different levels of participation and expertise. For information hazards such as CBRN risks, it may be useful to include subject-matter experts in red-teaming exercises in order to distinguish between cases in which expert knowledge is the key differentiating factor between dangerous and non-dangerous use of a foundation model. For particularly sensitive fields, such as nuclear security, this knowledge may be classified and/or heavily concentrated within the federal government. In contrast, assessing societal-level risks such as discriminatory output or systemic hallucinations should involve participation from relevant stakeholders, including members of potentially affected categories or communities.
    - A truly multi-stakeholder approach may be labor intensive and involve skilling up participants who may be unfamiliar with foundation models' capabilities and limitations, but such an investment may be essential to the success of this kind of red-teaming exercise.
- Data related to red-teaming, such as logs containing prompts and responses, should also be secured via appropriate access controls and cybersecurity best practices.

Criteria for Defining an Error, Incident, or Negative Impact

Relevance: NIST RFI Section 1 on *Developing Guidelines, Standards, and Best Practices for AI Safety and Security*; E.O. 14110 Sections 4.1(a)(i)(A) and (C)
Referenced Publications: Adding Structure to AI Harm: An Introduction to CSET's AI Harm Framework; Understanding AI Harms: An Overview

- CSET defines "AI harm" using four components: AI harm occurs when an affected entity experiences harm or potential for harm that is directly linked to the behavior of an AI system.
- Keeping these four components in mind, there are three routes by which harm can occur.
  - An AI system can unintentionally cause harm.
  - An AI system can intentionally cause harm.
  - An AI system itself can be harmed by entities, which include other AI systems or people.
- CSET's framework for AI harm covers common categories of harm, including harm to physical health or safety, financial loss, property damage, detrimental content, bias and differential treatment, and violation of privacy, human and civil rights, or democratic norms.
- AI harm can involve a single AI system and affected entity, but it can also involve multiple harmed entities, AI systems, or types of harm. An AI system could, for example, harm multiple entities, harm the same entity in multiple ways, or both.
- Notably, CSET's framework for AI harm distinguishes harm that actually occurred from harm that may occur. This enables tracking of realized harms, while also supporting analysis on potential harms.

Governance Policies and Technical Requirements

Relevance: NIST RFI Section 1 on *Developing Guidelines, Standards, and Best Practices for AI Safety and Security*; E.O. 14110 Sections 4.1(a)(i)(A) and (C)
Referenced Publications: AI Incident Collection: An Observational Study of the Great AI Experiment; Special Competitive Studies Project & Johns Hopkins University Applied Physics Laboratory: Framework for Identifying Highly Consequential AI Use Cases

- Quality AI incident collection requires clear goals that support actions, collaboration, analyzable and meaningful data, clear and specific requirements, infrastructure that is easy to use, updateable processes, adequate resourcing, and identified roles, responsibilities, and authorities. The United States will likely need multiple different AI incident reporting regimes to effectively mitigate risk from AI systems and balance out the limitations of any one reporting regime. These regimes include mandatory, voluntary, and citizen reporting regimes.
    - Mandatory reporting requires stakeholders to report incidents with specific information. Mandatory reporting can enforce consistency among reports but may impose a large administrative burden and lack the flexibility to accommodate changes in the nature of incidents.
    - Voluntary reporting gives stakeholders the option to report incidents, often with guidance on how to report specific information. Voluntary reporting can be more flexible and less resource-intensive than mandatory reporting, but may exclude important information that companies are reluctant to disclose.
    - Citizen reporting is conducted by stakeholders who serve as watchdogs. Citizen reporting can be spun up quickly and is likely to catch novel harms, but of the three reporting regimes, it is the most prone to inconsistent reports.
- Multiple sets of governance policies are needed for errors, incidents, and negative impacts (we will refer to these collectively as just “errors”). Developers and deployers need internal governance policies for tracing errors. Ideally, NIST would provide guidance on appropriate governance for AI errors. Not all of the items tracked internally would need to be reported to regulators or the

government. Instead, serious or novel errors would have mandatory reporting requirements. Regulators and the federal government should develop guidance and requirements for identifying, defining, and reporting serious events.

- When assessing the seriousness of errors, assessments should not be based solely on the intensity of harm resulting from errors but also upon the number of people who were exposed (or continue to be exposed) to the error. Without accounting for the number of people exposed to the error, a prolonged error that results in minor harms to large populations would typically not be considered serious, which may be problematic.
- The Special Competitive Studies Project and Johns Hopkins University Applied Physics Laboratory (see page 7) have proposed four factors for determining the severity of an error: scale, scope, disproportionality, and duration. CSET is in the process of determining if these factors can be applied to incidents in the AI Incident Database. Preliminary results suggest that they can, although the range of values for each factor will vary depending on the types of harm that result from an error (financial harm, physical harm, etc.). This implies that the severity levels of different harm types are not easily comparable.

AI Risk Management and Governance

Relevance: NIST RFI Section 3 on *Advance Responsible Global Technical Standards for AI Development*; E.O. 14110 Section 11(b)
Referenced Publications: Translating AI Risk Management Into Practice

- One strategy for managing the risk of an AI system is to consider risk that may arise at each stage of the system's lifecycle. Documentation that emphasizes the following items can help accomplish this:
    - Descriptions of all stakeholders that are not a decision-making authority, including their responsibilities:
        - If appropriate, clarify intended users of the system and their expectations; and,
        - Consider stakeholders that are not users or directly part of the AI lifecycle – for example, those indirectly affected by the AI system
    - Who has the decision-making authority to:
        - Sign off on tasks
        - Accept risk or a risk mitigation approach
    - The decision points at each stage of the AI system's lifecycle:
        - Decision authority and accountability at each decision point
        - Entrance and exit criteria at each decision point
        - Outputs and documentation at each decision point
    - The resources at each lifecycle stage, including
        - Workforce needs, including any special skill sets
        - Compute
        - Hardware
        - Test harnesses
        - Access to high-demand or rare resources
- For stakeholders who wish to implement the NIST AI RMF, this documentation can aid in the prioritization of high-level risk management actions within the AI RMF by linking actions to clearly defined risk management roles and timelines.

Strategies for Driving Adoption and Implementation of International Standards

Relevance: NIST RFI Section 3 on *Advance Responsible Global Technical Standards for AI Development*; E.O. 14110 Section 11(b)
Referenced Publications: Repurposing the Wheel: Lessons for AI Standards

- An independent body not involved in certification, such as a Federally Funded Research and Development Center, should conduct research into how to most effectively use third-party accreditation bodies to promote consistent AI standards implementation. As risks that characterize AI systems are still being understood, particular care should be taken to ensure that accreditation bodies have appropriate incentives and capabilities to evaluate the implementation of standards for AI systems. This study could support:
    - Implementation of the October 30, 2023 Executive Order on AI by informing best practices for ensuring that evaluations of vendor claims are performed in a consistent and fair manner; and/or,
    - Implementation of the Office of Management and Budget's draft guidance on Advancing Governance, Innovation, and Risk Management for Agency Use of AI by helping independent evaluation authorities standardize their review of agencies' AI systems.
- Professional organizations such as the Association for Computing Machinery should establish AI standards access funds, whistleblower protection programs, and reporting programs to gather anonymized information on AI risks from industry participants.
    - The cost of purchasing private standards or certification can be prohibitive for certain businesses. Professional organizations should establish AI standards access funds for small- to medium-sized businesses that cannot afford to access standards behind a paywall.
    - Professional organizations should establish whistleblower protection programs to ensure employees are not exposed to undue risk from reporting AI standards compliance violations.
    - Reports about AI risks submitted to professional organizations through information-gathering initiatives should not be traceable to individual

companies so that companies are willing to have their employees participate without reputational risk. The findings of such programs should be shared, at least in summary form, with industry and government stakeholders to inform risk mitigation measures and standards development. These findings could help identify best practices for developing, deploying, and using AI systems and point towards areas where stronger oversight is needed.

Potential Mechanisms for Promoting International Collaboration

Relevance: NIST RFI Section 3 on *Advance Responsible Global Technical Standards for AI Development*; E.O. 14110 Section 11(b)
Referenced Publications: Repurposing the Wheel: Lessons for AI Standards

- Building off of the G7 Hiroshima Process, the United States should commence discussions about creating the equivalent of a Financial Action Task Force (FATF) for AI.
    - The FATF is an intergovernmental body that sets international standards to counter global money laundering and terrorist financing. The FATF uses “black” and “grey” lists to identify countries with weak enforcement of standards, and then names these countries in public documents three times a year. The black list serves as a call to action for FATF members to apply enhanced due diligence, and sometimes countermeasures, against those countries considered high risk. The grey list names countries subject to increased monitoring as they address deficiencies.
    - A structure akin to the FATF that leverages grey and black lists for AI would place international pressure on countries and other entities to create and implement guardrails for the safe development, deployment, and use of AI systems. Grey lists should include companies and countries that have engaged in questionable conduct around AI. Black lists should include companies and countries that have a documented history of unsafe behavior and show resistance to changing their behavior.
- NIST should create an online portal to ensure technical developments relevant to standards are captured and publicized.
    - The portal could be housed within the existing Trustworthy and Responsible AI Resource Center as an unofficial addendum to the AI RMF, where industry stakeholders can provide real-time updates of AI advancements, such as substantial increases in capabilities of AI systems or decreases in resources required for given capabilities.
    - While NIST should provide oversight of submissions for quality purposes, this portal should serve as an accessible mechanism for entities that may encounter new problems outside the scope of the existing guidance.

  - Since the portal would provide an opportunity for prototyping AI standardization language, it would serve as a resource for NIST when the time comes to formally update the AI RMF, which would potentially shorten the revision timeline.
- Standard-setting bodies should host biannual summits to coordinate on standards interoperability and efficacy.
  - By harmonizing review processes during biannual gatherings, regulators may be able to create multiple layers of protection against harms from general-purpose AI systems.
  - Hosting regular gatherings will sustain progress on standards for AI evaluation metrics and safety testing, supporting the goals of concordant events such as future international AI safety summits, like the follow-up summits to the UK’s November 2023 Bletchley Summit planned for May and November of 2024.
- NIST should support the development of testbeds to monitor AI standards for effectiveness.
  - The CHIPS and Science Act of 2022 authorized NIST to create testbeds for “safe and trustworthy artificial intelligence and data science.” NIST should support the development of these testbeds and adapt them to function like regulatory sandboxes, but for voluntary standards.
  - Organizations that build AI systems could determine whether their systems adequately implement standards in the testbed with help from subject matter experts. In turn, NIST could proactively monitor the ease with which organizations adhere to standards, assess if technology has surpassed the scope of existing standards, and vet third-party proposals for additional metrics, benchmarks, and standards.
  - Testbeds for standards implementation would complement the testbeds created by the Secretary of Energy and NSF under the October 30, 2023 Executive Order to advance the safe development of AI technologies.