Response to NIST RFI Related to the AI Executive Order

1 February 2024

Elham Tabassi, Chief of Staff, 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

Subject: Request for Information - NIST's Assignments under Executive Order 14110 on Safe, Secure, and Trustworthy Development and Use of AI

Thank you for the opportunity to submit 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.

We are a technical research non-profit that works to reduce risks associated with AI.

In our response, we focus primarily on guidance and benchmarks for evaluating and auditing AI capabilities and guidelines for AI red-teaming tests, particularly in the context of dual-use foundation models (E.O. Sec. 3. (k)). We also comment on some best practices and supporting resources that could be incorporated into a companion resource to the AI Risk Management Framework.

We start with a summary of our recommendations and then provide supporting detail and references to current best practices.

Thank you again for this opportunity to comment. We look forward to continuing to support NIST's crucial work on AI standards.

Yours sincerely,

Dan Hendrycks
Director, Center for AI Safety

Recommendations

AI RMF Companion (1.a.1)

1. Structure: NIST's guidance should **distinguish between different categories of generative AI systems based on the level of risk they present**. For example, the Partnership on AI's Guidance for Safe Foundation Model Deployment distinguishes between "Specialized Narrow Purpose", "Advanced Narrow and General Purpose" and "Paradigm-shifting or Frontier" systems.
2. Principles for models with dual-use risks: Guidance on risk management for generative AI systems should reflect well-established principles in safety engineering and cybersecurity such as:
    a. **Defense-in-depth**: layering multiple independent risk management techniques
    b. **Safe by design**: addressing safety considerations early in the model development lifecycle
3. Best practices for models with dual-use risks: **NIST's guidance should provide concrete examples of responsible risk management practices for AI developers developing systems in the highest risk tier**. Specific examples of relevant risk management practices from the existing literature include:
    a. **Scanning for novel and emerging risks** on a regular basis
    b. **Responsible iteration**, including pre-development risk assessments, gradual scaling and ongoing testing during development, and staged release
    c. Developing a **safety case** providing evidence that a model is safe to develop
    d. Setting up **credible response plans** for mitigating risks that are only identified after a model is released, and running drills to test these
    e. **Cybersecurity and information security** measures that are proportionate to the value of the AI system to potential attackers (up to and including nation-states)
4. Removing hazardous capabilities: NIST should provide guidance on **techniques to reduce hazardous capabilities of dual-use foundation, such as exacerbating CBRN risks or facilitating cyber-attacks**. This would complement discussion of evaluation of hazardous capabilities, which is insufficient on its own to equip developers to manage these risks adequately. Guidance on reducing these capabilities should **discuss a range of techniques that can enable a "defense-in-depth" approach when dealing with poorly understood systems**, such as filtering of data, adversarial training, blocklists, prompt transformations, use of input and output classifiers, and **machine unlearning techniques that aim to reduce a model's hazardous capabilities.** We believe significant progress in machine unlearning is possible, and **CAIS will shortly publish relevant results from our research on improving unlearning techniques, applied specifically to CBRN-related knowledge**.

Guidance and benchmarks for evaluating and auditing AI capabilities (1.a.2)

5. Threat identification and prioritization: NIST should provide guidance on **threat modeling to support AI developers in prioritizing which threats to focus their evaluation and red-teaming efforts on**. Without sufficiently structured and well-resourced efforts to identify and prioritize risks, there is a risk that evaluations will

focus on well-understood risk areas and ignore other risks, resulting in blind spots and a false sense of security.

6. Benchmarks to test for hazardous capabilities: NIST's guidance should promote the use of systematic, quantitative evaluations to test for hazardous capabilities, in addition to red-teaming by subject matter experts or others. However, this will require progress on creation and validation of relevant evaluations, as there are currently few publicly available tools to support evaluations for many categories where AI systems could exacerbate risks (such as CBRN risks). For example, CAIS has recently created a **benchmark for evaluating hazardous knowledge of biology, chemistry and cybersecurity in text-based generative AI models**, which can be combined with the machine unlearning techniques mentioned above to reduce hazardous capabilities. CAIS expects to make this available to other researchers in a responsible manner, while avoiding proliferation of information that could be used to cause significant harm.

Guidelines for AI red-teaming tests (1.b)

7. Taxonomy of red-teaming: NIST's guidance on red-teaming should **distinguish between different potential models for red-teaming and define these clearly**. AI developers and policy-makers have so far used this term to refer to a range of practices for assurance, ranging from small-scale internal stress testing to crowdsourcing vulnerabilities from members of the public, as well as automated adversarial testing approaches.
8. Assurance against Advanced Persistent Threats or similar actors: NIST's guidance should discuss how red-teaming exercises can **realistically simulate well-resourced and highly motivated adversaries** whose approaches are creative and evolve over time in response to defenses by AI developers. This is particularly important in the context of dual-use foundation models that may be an attractive target for nation-states or other persistent adversaries.

AI RMF - companion document for generative AI (1.a.1)

Structure

- NIST's guidelines for generative AI should avoid a one-size-fits-all approach to risk management. The Executive Order highlights dual-use foundation models with "at least tens of billions of parameters" as particularly likely to pose serious risks to national security (E.O. Sec. 3. (k)). More broadly, risk management is more challenging for systems that are highly general in the range of tasks they can accomplish, modalities they can handle, etc., and for systems that are novel, for example, in terms of the amount of computation or techniques used to train them (Weidinger et al. 2022, Wei et al. 2022). Models operating across multiple modalities or that can power "AI agents" carrying out a sequence of actions without immediate human oversight likely also require a different approach due to the increased risks they may present and the more limited research on them (Hendrycks et al. 2023, Shavit et al. 2023). NIST should also consider

risks from narrow models with high-risk applications in specific sectors such as biology. For example, models trained on protein or DNA sequence data could present dual-use risks, and are seeing a rapid growth in the computation used to train them which could enhance their capabilities and exacerbate these risks (Maug et al. 2024).
- NIST should consider adopting a tiered approach in its guidelines with more stringent risk management measures for AI systems that are less well-understood and more likely to present severe risks, or alternatively publish separate guidance focused on these systems. Such an approach would be in line with the recent Executive Order, which identified dual-use foundation models as presenting particular risks, and the Partnership on AI's recent Guidance for Safe Foundation Model Deployment.

Principles for models with dual-use risks

- Guidelines should aim for a "defense-in-depth" approach to risk management, particularly for AI systems whose risks and capabilities are poorly understood due to their novelty, generality or other features (Ee et al. 2023). The core idea of defense in depth is that it is unlikely that any one layer of defense is foolproof; we are usually engaging in risk reduction, not risk elimination. For example, in order to reduce risks that models will assist with the creation of biological weapons, developers could be encouraged to layer a range of measures such as filtering of data, machine unlearning, adversarial training, blocklists, prompt transformations and use of input and output classifiers.
- Guidelines should also promote a "safe by design" approach where safety assurance and guardrails are embedded early in the development cycle rather than being added at the end (WEF AI Governance Alliance, 2023). An example of this approach would be encouraging developers to use data curation and machine unlearning approaches to remove as far as possible hazardous capabilities of dual-use foundation models. This contrasts with an approach to risk mitigation based solely on adding restrictions to existing models to prevent users from taking advantage of hazardous capabilities. Such restrictions have been shown to be fragile and easy to circumvent in a variety of ways (Wei et al. 2023, Gade et al. 2023).

Best practices for models with dual-use risks

- NIST's guidance on risk management for generative AI systems should discuss best practices for general-purpose AI systems that may present dual-use risks. NIST draw on existing recommendations on this topic from the Partnership on AI, researchers at UC Berkeley and the UK AI Safety Summit, among other resources.
- We highlight the following practices for managing risks from foundation models, many of which are also included in the Partnership on AI's Guidance for Safe Foundation Model

Deployment. These would help developers of dual-use foundation models to operationalise many of the principles included in the AI RMF:

a. **Scan for novel or emerging risks**: AI developers should conduct ongoing analysis to identify unknown risks or threat models, particularly risks not adequately addressed by their existing risk management practices (builds upon RMF Measure 3.1 and 3.2; see also Govern 4.2, 5.1 and Map 3.2, 5.1)
b. **Responsible iteration**: When developing generative AI systems that are at or beyond the current research frontier and may pose dual-use or other severe risks, AI labs should adopt best practices such as **pre-development risk assessments, gradual scaling and ongoing testing during development, and staged model release** (builds upon RMF Govern 4.3, 5.1 and Measure 2.3, 2.5, 2.6)
c. **"Pre-Systems Card' and safety case**: AI developers building models beyond the current research frontier should disclose their planned evaluation and risk management procedures for frontier models before development. They should prepare evidence-based, affirmative arguments explaining why the model is safe enough to develop. One form this could take is a written "safety case", similar to those used in aviation and other industries (builds upon RMF Map 1.5, 2.3, 5.1)
d. **Monitoring and response plans**: AI developers should monitor systems for misuse or unintended uses, as well as other risks. They should put in place plans for how they respond to security incidents such as cyberattacks and safety incidents such as misuse of AI systems (builds upon RMF Manage 2.3, 4.1 and Govern 6.2). They should conduct drills to test the effectiveness of these plans and address any material identified issues.
e. **Cybersecurity**: AI developers should rigorously test and address security vulnerabilities in frontier models, considering risks such as prompt insertion attacks, training data extraction, backdoors, adversarial examples, data poisoning and model exfiltration (builds upon RMF Measure 2.7).
f. **Information security**: State, industry, and criminal actors are motivated to steal model weights and research IP and AI labs should take measures in proportion to the value and risk level of their IP. Eventually, this may require matching or exceeding the information security of military or security agencies, since attackers may include nation-states. Information security measures include not only defending against external cyber-attacks, but also implementing measures to address insider threats, physical infiltration of AI developer premises, or other threats (builds upon RMF Measure 2.7).

By providing a range of concrete best practices for safe AI development across the full model lifecycle, NIST's guidelines can help AI developers to mitigate risks more effectively. When developing models whose risks are poorly understood, developers should build in risk management throughout the model lifecycle and adopt a defense-in-depth approach that layers many risk management techniques.

Removing hazardous capabilities

Removing hazardous capabilities from the model is an example of a “secure by design” approach that builds in security as early as possible in the model lifecycle. One set of techniques to achieve this objective involves filtering training data so as to exclude relevant knowledge from the training data. Another set of approaches involve model unlearning, which aims to remove certain hazardous or sensitive knowledge from existing pre-trained or fine-tuned models. This is a promising additional line of defense to help with mitigating risks relating to information hazards such as facilitating the development of biological or chemical weapons, as well as leaking other private or sensitive information contained in training data. CAIS researchers have developed techniques (to be published shortly) that degrade the ability of models to successfully answer questions on CBRN-related knowledge, thereby reducing their ability to output knowledge that could be used to cause harm. The associated dataset used for this is further described below.

Guidance and benchmarks for evaluating and auditing AI capabilities (1.a.2)

Threat identification and prioritization

NIST should provide guidance on techniques for AI developers to identify novel risks and threat models on an ongoing basis, and on how to prioritize risks for further investigation and assessment. This is important given that generative AI systems may exhibit emergent capabilities, that our techniques for evaluation are highly immature and not exhaustive, and that systems may interact with society in complex and unanticipated ways. Therefore developers need to be equipped to identify and mitigate risks that have not yet been identified internally or in external guidelines and regulations. Relevant techniques identified in the literature (e.g. Hicks et al. 2023) include:

- Forecasts, simulations and other futures methods
- Engagement with affected individuals and communities
- External data or model audits
- Pilot studies in sandboxes or other lower-risk environments
- Ongoing adversarial testing/ red-teaming
- Continuous monitoring of use

Guidance on threat modeling would support AI developers in prioritizing which threats to focus their red-teaming efforts on. For example, Anthropic’s Responsible Scaling Policy and findings from its work on red-teaming emphasize the role of threat models as a starting point for identifying which capabilities need to be assessed and how evaluations should be designed. Similarly, OpenAI’s Preparedness Framework highlights seeking out “unknown unknowns” as a priority in order to identify categories of risk that have not yet been identified.

Without clear processes for specifying and prioritizing threat models, there is a risk that evaluations will prioritize known risks and fail to explore other areas (Feffer et al., 2024). These

biases could then become self-reinforcing. For example, the selection of subject matter experts used in red-teaming and evaluation could influence the types of risks that are evaluated and the methodologies used. Similarly, crowd-sourced teams of red-teamers that are under time constraints may default to focusing on well-understood risk areas where they can quickly find successful attacks and fail to perform testing in areas that are more complex and time-consuming.

Benchmarks to test harmful capabilities

There are currently few or no publicly available benchmarks to support evaluation of hazardous capabilities such as AI systems' ability to exacerbate CBRN risks or facilitate cyber-attacks. To identify gaps in existing benchmarks and evaluations, NIST can take advantage of existing repositories and evaluation suites such as Google DeepMind's Sociotechnical Safety Evaluation Repository, SafetyPrompts, DecodingTrust and Cataloguing LLM Evaluations.

Researchers at the Center for AI Safety have developed tools for the evaluation of hazardous capabilities in text-based systems, including a benchmark to assess the ability of a model to output harmful content associated with biological, cybersecurity, or chemical risks. This is based on a large dataset of questions and answers collected from specialists in these fields. The data aims to measure knowledge that is a precursor to or correlated with hazardous capabilities, but the dataset itself does not include highly sensitive knowledge. Examples of knowledge measured would include certain types of virology knowledge for biosecurity, or penetration knowledge for cybersecurity. We expect to publish this dataset in the near future and hope that this benchmark will enable more systematic, automated and continuous evaluation of the risks posed by language models in these domains.

Guidelines for AI red-teaming tests (1.b)

Taxonomy of red-teaming

NIST's guidance should provide a taxonomy of approaches to red-teaming with clear definitions. Red-teaming can have a number of meanings, which has created a lack of clarity in discussions around its role. A recent review of the literature on AI red-teaming by Feffer et al. (2024) highlights divergences between previous red-teaming exercises in terms of:

- **Criteria of evaluation**: specific threat models or risks assessed; choice between targeted assessment of specific risks vs. broad and open-ended exploration of model capabilities
- **Process**: single-round vs. iterative red-teaming; time allocated to red-teamers
- **Techniques**: brute force generation of concerning outputs using humans or language models, algorithmic search, targeted attacks on specific system components
- **Assumptions around adversary capabilities**, which are implicitly reflected in the level of resources and time given to red-teamers to simulate adversaries

- **Team composition**: this can include selected subject matter experts in relevant areas, crowd-sourced workers from online platforms, competition participants or language models used to red-team themselves
- **Disclosure of costs, findings and response** (i.e. mitigations)

NIST's red-teaming guidance should include standards around documentation and disclosure of findings and risk mitigations. Reporting by developers that shares the risk areas assessed (or deliberately not assessed), the findings and the measures taken in response to these would make it easier to assess the effectiveness of red-teaming exercises and remaining vulnerabilities that users, regulators and other stakeholders should be aware of. There are several reported examples where AI developers have been notified of vulnerabilities in their models by researchers or others (for example, Wei et al. 2023 note that they reached out to OpenAI and Anthropic to share their findings), but developers have been unable or unwilling to mitigate some of these issues.

Assurance against Advanced Persistent Threats or similar actors

NIST's guidance should discuss how red-teaming exercises can realistically simulate well-resourced and highly motivated adversaries whose approaches are creative and evolve over time in response to defenses by AI developers. One recent attempt at this is Mouton et al. (2024)'s research on the potential of Large Language Models to increase the risk of large-scale biological attacks.

Standardization of red-teaming exercises brings some challenges, as imposing a prescriptive list of tests to be conducted could prevent red-teamers from surfacing previously unidentified threats. Given the inherent uncertainty in predicting future threats, red-teaming guidelines need to leave room for creativity on the part of red-teamers in order to be useful in pre-empting real-world risks. This is particularly true when conducting red-teaming exercises on dual-use foundation models, which may be targeted by state or non-state actors that are highly persistent.

Appendix - further detail on red-teaming best practices

CAIS researchers have first-hand experience with red-teaming major generative AI systems and have worked with both OpenAI and Meta on red-teaming several systems. In this appendix, we provide further details on red-teaming processes, and describe a few best practices identified based on our experience and the available literature. Feffer et al. (2024) provides a review of the existing literature on AI red-teaming. Ganguli et al. (2022) describes an early red-teaming process used by Anthropic for their Claude models. Other examples that offer some detail include Achiam et al. (2023) and Touvron et al. (2023).

Red-teaming processes

Red-teaming of general-purpose models such as GPT-4 or Llama-2 was a multi-month process which involved red-teamers testing the model at several stages in its development. For example, in GPT-4's case, some red-teamers were given access to a model which had undergone an initial round of RLHF but before further safety measures had been applied. They were then asked to test the model again after additional safety measures have been added, such as fine-tuning, adversarial training, blocklists, prompt transformations, use of input and output classifiers to identify prompts that should be rejected or outputs that should not be provided (see the DALL-E 3 System Card for examples of mitigations used by OpenAI).

As AI developers have gained more experience with red-teaming, they increasingly provide customized user interfaces that allow red-teamers to easily interact with the model, flag interactions for review and provide additional information (types of harm or policies broken, severity of issue) within a single interface. Interactions that have been flagged are typically stored in a database for later review. In some cases this dataset may be used for automated tests at a later stage once the system has been updated with risk mitigations. At this stage, all the prompts identified as causing issues during red-teaming are run through the model again to check if outputs are still problematic. There are no current standards around sharing this information: Anthropic released a dataset based on some red-teaming exercises (Ganguli et al., 2022) whereas other developers have provided more limited information on the issues identified. As discussed under our recommendations on red-teaming, transparent reporting and structured processes for sharing information about vulnerabilities would be valuable.

Best practices

There can be many cases where it is unclear how severe the potential harm from an output is. In the absence of guidance provided by the AI developer, there is a risk that red-teamers apply inconsistent standards or do not set an appropriate threshold for when to flag issues. It is therefore useful for AI developers to provide guidelines for red-teamers with examples of concerning prompts and outputs across various categories of harms or risks. These guidelines should not aim to be exhaustive and should still leave scope for red-teamers to flag issues that fall outside their taxonomy of risks.

The expertise required among red-teaming participants will depend on the threat models and major risks that have been previously identified. For the most capable and general systems, a broader range of threat models should be tested, given the higher degree of uncertainty around potential capabilities. The GPT-4 Technical Report (Achiam et al. 2023), OpenAI Preparedness Framework and Red Teaming Network (2023) and Llama-2 release paper (Touvron et al. 2023) provide examples of domains from which experts have been recruited for red-teaming. To test model capabilities relating to national security risks, there might be a need to work with experts that have security clearances and access to classified information, creating legal challenges (Anthropic, 2023).

Red teaming exercises for dual-use foundation models should include people with experience in prompt engineering or white-hat hacking that have a track record of creatively finding ways of circumventing safeguards, potentially working in collaboration with domain experts. A primer on

prompting techniques should also be provided to all red-teaming participants. This helps to ensure that experts are effectively testing the upper limits of the information that can be extracted from models, rather than being held back by limited creativity around prompts or other techniques.

In addition to experts, it is important to include red-teamers that are representative of the expected user base of the system. Due to their focus on risks related to the area of expertise, domain experts may overlook ways in which systems may produce undesirable outputs in the course of ordinary use by non-specialists. For example, when dealing with generative AI systems that work with inputs or outputs in multiple languages, it is important to ensure that red-teaming and other testing is conducted in multiple languages. There is evidence that it is often possible to “jailbreak” generative AI systems by using languages other than English, particularly low-resource languages, as measures taken to prevent harmful outputs in English may not transfer to other languages (Yong et al. 2023). Alternatively, some generative AI developers choose to restrict their models to English in order to avoid these vulnerabilities.

References

Achiam, Josh et al.(2023), GPT-4 Technical Report

AI Verify Foundation (2023), Cataloguing LLM Evaluations

Anthropic (2023), Challenges in evaluating AI systems

Anthropic (2023), Responsible Scaling Policy

Barrett, Anthony et al. (2023), AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models: Second Full Draft

Ee, Shaun et al. (2023), Adapting cybersecurity frameworks to manage frontier AI risks: a defense-in-depth approach

Feffer, Michael et al. (2024), Red-Teaming for Generative AI: Silver Bullet or Security Theater?

Gade, Pranav et al. (2023), BadLlama: cheaply removing safety fine-tuning from Llama 2-Chat 13B

Ganguli, Deep et al. (2022), Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned

Hendrycks, Dan et al. (2023), Overview of Catastrophic AI Risks

Hicks, Marie-Laure et al. (2023), Exploring red teaming to identify new and emerging risks from AI foundation models

Maug, Nicole et al. (2024), Biological Sequence Models in the Context of the AI Directives

Mouton, Christopher et al. (2024) The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study

OpenAI (2023), DALL-E 3 System Card

OpenAI (2023), Preparedness Framework

OpenAI (2023), OpenAI Red Teaming Network

Partnership on AI (2023), Deployment Guidance for Foundation Model Safety

Shavit, Yonadav et al. (2023), Practices for Governing Agentic AI Systems

Touvron, Henri et al. (2023), Llama 2: Open foundation and fine-tuned chat models

UK Government (2023), Emerging Processes for Frontier AI Safety

Wang, Boxin et al. (2023), DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models

Wei, Alexander et al. (2023), Jailbroken: How Does LLM Safety Training Fail?

Wei, Jason et al. (2022) Emergent Abilities of Large Language Models

Weidinger, Laura et al. (2022), Taxonomy of Risks posed by Language Models. In 2022 ACM Conference on Fairness, Accountability, and Transparency

Weidinger, Laura et al. (2023), Sociotechnical Safety Evaluation of Generative AI Systems

World Economic Forum (2023), AI Governance Alliance Briefing Paper Series January 2024 - Presidio AI Framework: Towards Safe Generative AI Models

Yong, Zheng-Xin et al. (2023), Low-Resource Languages Jailbreak GPT-4