Response to Request for Information (RFI NIST-2023-0009-0001) 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)

Organization: Future of Life Institute

Point of Contact: Hamza Tariq Chaudhry, US Policy Specialist. hamza@futureoflife.org

About the Organization

The Future of Life Institute (FLI) is an independent nonprofit organization with the goal of reducing large-scale risks and steering transformative technologies to benefit humanity, with a particular focus on artificial intelligence. Since its founding, FLI has taken a leading role in advancing key disciplines such as AI governance, AI safety, and trustworthy and responsible AI, and is widely considered to be among the first civil society actors focused on these issues. FLI was responsible for convening the first major conference on AI safety in Puerto Rico in 2015, and for publishing the Asilomar AI principles, one of the earliest and most influential frameworks for the governance of artificial intelligence, in 2017. FLI is the UN Secretary General's designated civil society organization for recommendations on the governance of AI and has played a central role in deliberations regarding the EU AI Act's treatment of risks from AI. FLI has also worked actively within the United States on legislation and executive directives concerning AI. Members of our team have contributed extensive feedback to the development of the NIST AI Risk Management Framework, testified at Senate AI Insight Forums, participated in the UK AI Summit, and connected leading experts in the policy and technical domains to policymakers across the US government.

EXECUTIVE SUMMARY

We would like to thank the National Institute of Standards and Technology (NIST) for the opportunity to provide comments regarding 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). The Future of Life Institute (FLI) has a long-standing tradition of work on AI governance to mitigate the risks and maximize the benefits of artificial intelligence. In NIST's implementation of the Executive Order 13960 on "Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government" (EO), we recommend consideration of the following:

- **Military and national security AI use-cases should not be exempt from guidance.** Military and national security AI use-cases, despite exemptions in the Executive Order, should not be beyond NIST's guidance, considering their potential for serious harm. Given their previous work, NIST is well-positioned to incorporate standards for military and national security into their guidelines and standards.
- **A companion resource to the AI Risk Management Framework (RMF) should explicitly characterize minimum criteria for unacceptable risks.** NIST should guide the establishment of tolerable risk thresholds. These thresholds should include guidelines on determining unacceptable risk and outline enforcement mechanisms and incentives to encourage compliance.
- **Responsibility for managing risks inherent to AI systems should fall primarily on developers.** The NIST companion resource for generative AI should define roles for developers, deployers, and end-users in the assessment process. Developers, end-users, and deployers should all work to mitigate risk, but the responsibility of developers is paramount, considering their central role in ensuring the safety of systems.
- **Dual-use foundation models developed by AI companies should require external red-teaming.** External red-teams are essential for encouraging comprehensive and unbiased assessments of AI models. NIST should establish standards to ensure that external auditors remain independent and aligned with best practices.
- **The NIST companion resource for generative AI should include specific guidance for AI models with widely available model weights.** Safeguards designed to mitigate risks from dual-use foundations models with widely available weights can be easily removed and require specific standards to ensure security.
- **Embedded provenance on synthetic content should include developer and model information.** Including information on synthetic content about the developer and system of origin would better inform consumers and incentivize developers to prioritize safety from the design phase.

- **NIST should adopt a less restrictive definition of “dual-use foundation models.”** Switching from a restrictive definition (using 'and) to a more expansive definition (using 'or') as stated in the EO would enable NIST to bring all models of concern within its purview.

We look forward to continuing this correspondence and to serve as a resource for NIST efforts pertaining to AI in the months and years to come.

RECOMMENDATIONS

1. Military and national security use-cases

Standards for national security and military are not beyond the remit of NIST. AI systems intended for use in national security and military applications present some of the greatest potential for catastrophic risk due to their intended use in critical, often life-or-death circumstances. While the EO exempts national security and military AI from most of its provisions, NIST has previously established standards[1] related to national security, including standards for chemical, biological, radiological, nuclear, explosive (CBRNE) detection, personal protective equipment (PPE), and physical infrastructure resilience and security. Given this precedent, NIST can and should update the AI RMF and future companion pieces to include standards applicable to national security and military uses of AI. Specifically, NIST can play a vital role in mitigating risks presented by these systems by, *inter alia*, working with the Defense Technology Security Administration (DTSA) and the Office of Science and Technology to instate standards for procurement, development and deployment of AI technologies. Considering the sizable impact malfunction, misuse, or malicious use of military or national security AI systems could entail, such standards should be at least as rigorous in assessing and mitigating potential risks as those developed for civilian AI applications.

2. Addressing AI RMF gaps

NIST should provide guidance on identifying unacceptable risks. The AI risk management framework lacks guidance on tolerable risk thresholds. As a result, developers of potentially dangerous AI systems can remain in compliance with the AI RMF despite failure to meaningfully mitigate substantial risks, so long as they document identification of the risk and determine that risk to be acceptable to them. Accordingly, companies can interpret risk solely in terms of their interests - tolerable risk may be construed as risks that are tolerable for the developer, even if those risks are unacceptable to other affected parties. The ability to make internal determinations of tolerable risk without a framework for evaluating externalities overlooks the potential impact on government, individuals, and society. NIST should revise the AI RMF, introducing criteria for determining tolerable risk thresholds. This revision should incorporate evaluations of risk to individuals, communities, and society at each stage of the assessment process, and these revisions should be applied to all relevant companion resources.

---

[1] *Public Safety - National Security Standards.* National Institute of Standards and Technology. Accessed at: https://www.nist.gov/national-security-standards

Enforcement mechanisms and structural incentives are necessary. While industries may voluntarily adopt NIST standards, we cannot rely on AI companies to continue to self-regulate. The significance of these standards warrants explicit commitment through structured incentives and enforcement measures. To encourage the adoption of these standards, NIST should offer independent evaluation of systems and practices for compliance with their framework, provide feedback, and provide compliant parties with a certificate of accreditation that can demonstrate good faith and strengthen credibility with the public and other stakeholders.

Guidelines must set clear red-lines to halt or remediate projects. NIST should internally define minimum red-lines and encourage AI companies to predetermine additional red-lines for each assessment. Failure to stay within these limits should prevent the project from progressing or mandate remediation. Red-lines should encompass material risks of catastrophic harm and significant risks related to the ease and scale of misinformation, disinformation, fraud, and objectionable content like child sexual abuse material and defamatory media. Such predetermined, explicit thresholds for halting a project or taking remediation efforts will prevent movement of safety and ethical goalposts in the face of potential profits by companies, increasing the practical impact of the AI RMF's extensive guidance on assessment of risk.

3. AI developer responsibility

The NIST companion resource for generative AI should define clear roles for developers, deployers, and end-users in the assessment process. All of these parties should take steps to mitigate risks to the extent possible, but the role of the developer in proactively identifying, addressing, and continuously monitoring potential risks throughout the lifecycle of the AI system is paramount. This should include (but is not limited to) implementing robust risk mitigation strategies, regularly updating the system to address new vulnerabilities, and transparently communicating with deployers and end-users about the limitation and safe usage guidelines of the system.

Compared to downstream entities, developers have the most comprehensive understanding of how a system was trained, its behavior, implemented safeguards, architectural details, and potential vulnerabilities. This information is often withheld from the public for security or intellectual property reasons, significantly limiting the ability of deployers and end-users to understand the risks these systems may present. For this reason, deployers and end-users cannot be reasonably expected to anticipate, mitigate, or compensate harms to the extent that developers can.

Deployers implementing safety and security by design, and thus mitigating risks at the outset

prior to distribution, is more cost-effective, as the responsibility for the most intensive assessment and risk mitigation falls primarily on the handful of major companies developing advanced systems, rather than imposing these requirements on the more numerous, often resource-limited deployers. This upstream-approach to risk-mitigation also simplifies oversight, as monitoring a smaller group of developers is more manageable than overseeing the larger population of deployers and end-users. Furthermore, the ability of generative AI to trivialize and scale the proliferation of content makes dealing with the issue primarily at the level of the end user infeasible and may also necessitate more privacy-invasive surveillance to implement effectively.

Developer responsibility does not fully exempt deployers or end-users from liability in cases of intentional misuse or harmful modifications of the system. A framework including strict, joint and several liability, which holds all parties in the value chain accountable within their respective liability scopes, is appropriate. Failure by a developer to design a system with sufficient safeguards that cannot be easily circumvented should be considered akin to producing and distributing an inherently unsafe or defective product.

4. External red-teaming of dual-use foundation models

External red-teaming should be considered a best practice for AI safety. While many AI developers currently hire external teams with specialized knowledge to test their products, relying solely on developers to select these teams is insufficient due to inadequate standardization, conflicts of interest, and lack of expertise.

Ideally, the government would establish the capacity to serve in this role. However, in situations where government-led red-teaming is not feasible, alternative mechanisms must be in place. **NIST should move to establish a criteria to assess external auditors for their expertise and independence.**[2] These mechanisms could be implemented as an official certification displayed on the product's website, signifying that the model has passed testing by an approved entity. This approach not only enhances safety but also fosters transparency and public trust.

Ensuring comprehensive safety assessments requires red-teams to have access to the exact model intended for deployment, along with detailed information on implemented safeguards and internal red-teaming results. External testers are typically given "black-box" access to AI models

[2] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. 2022. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22). Association for Computing Machinery, New York, NY, USA, 557–571. https://doi.org/10.1145/3514094.3534181

via API access.[3] This approach limits their testing abilities to prompting the system and observing its outputs. While this is a necessary part of the assessment process, it is not sufficient and has shown to be unreliable in various ways.[4] Conversely, structured access provides testers with information that allows them to execute stronger, more comprehensive adversarial attacks. [5] Many companies oppose providing complete access to their models due to concerns about intellectual property and security leaks. To mitigate these concerns, we recommend that NIST establish physical and contractual standards and protocols to enable secure model access such as on-site testing environments and nondisclosure agreements. To ensure that external auditors are conducting tests in accordance with these standards and practices, these should be conducted by the government or other approved entities.

Red-teams should be afforded ample time, resources, and access for comprehensive testing. A multi-stage process including data, pre-training, model, system, deployment, and post-deployment phases is needed. Access to training data, for example, could foster transparency and enable pathways for the enforcement of copyright law. Furthermore, developers should be encouraged to proactively engage with deployers to understand the use-cases of their products and inform external auditors so that they may tailor their testing strategies effectively.

Finally, AI companies should be encouraged to establish mechanisms for the continuous identification and reporting of vulnerabilities post-deployment. Many companies have created pipelines for these processes.[6][7] NIST should consider providing guidelines to encourage consistency and standardization.

5. Safety limitations of AI models with widely available model weights

The NIST companion resource on generative AI should include recommendations on evaluating

---

[3] METR. (March 17, 2023). *Update on ARC's recent eval efforts.* Model Evaluation and Threat Research. Accessed at: https://metr.org/blog/2023-03-18-update-on-recent-evals/

[4] Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., ...and Hadfield-Menell, D. (2024). *Black-Box Access is Insufficient for Rigorous AI Audits*. arXiv preprint arXiv:2401.14446.

[5] Bucknall, B. S., and Trager, R. F. (2023). *Structured Access For Third-party Research On Frontier ai models: investigating researchers model access requirements.* Oxford Martin School AI Governance Initiative. Accessed at: https://www.oxfordmartin.ox.ac.uk/publications/structured-access-for-third-party-research-on-frontier-ai-models-investigating-researchers-model-access-requirements/

[6] Company Announcement. (July, 2023). *Frontier Threats Red Teaming for AI Safety.* Anthropic. Accessed at: https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety

[7] Blog. *OpenAI Red Teaming Network.* OpenAI. Accessed at: https://openai.com/blog/red-teaming-network

the risks of releasing models with widely available model weights. **With current technologies and architectures, removing safeguards from AI models with widely available model weights through fine-tuning is relatively trivial.**[8] This makes it intractable to set or enforce guidelines for developers who build on open-source models. This ease of removal has enabled the proliferation of harmful synthetic materials.[9]

6. Inclusion of developer information in synthetic content

Embedded information on synthetic content should include information about the developer and system of origin. Much attention has been paid in recent months to the potential for synthetic content to contribute to the spread of mis- and disinformation and non-consensual sexual imagery. The proliferation of synthetic content also carries significant national security risks, including the use of synthetic blackmail or spearphishing against high-ranking officials and the creation of fake intelligence, which could introduce serious vulnerabilities. Some generative AI systems may lack sufficient safeguards, making them more prone to these malicious uses, but detecting these vulnerabilities and holding their developers accountable for rectifying them is at present extremely challenging.

Labeling and watermarking techniques have been proposed as one possible method for verifying the authenticity or synthetic nature of content, and Section 4.5(a) of the EO tasks the Department of Commerce with developing or identifying existing tools, standards, methods, practices, and techniques for detecting, labeling, and authenticating synthetic content. We recommend that standards for watermarking or other embedded information should include information detailing the developer and system of origin. Such measures would incentivize developers to prioritize safety from the design phase, facilitate identification of systems especially vulnerable to creation of untoward content, and streamline the identification and tracking of problematic synthetic content back to its creators to impose liability for harms where appropriate. Given the stakes of the issues raised by synthetic content, the emphasis on safety and accountability should take precedent over concerns about the economic feasibility of implementation. That said, any additional economic burden for embedding system and developer of origin information would likely be negligible relative to embedding information relating to the authenticity of the content

[8] Qi, X., Zeng, Y., Xie, T., Chen, P. Y., Jia, R., Mittal, P., & Henderson, P. (2023). *Fine-tuning aligned language models compromises safety, even when users do not intend to!.* arXiv preprint arXiv:2310.03693. Accessed at: https://arxiv.org/abs/2310.03693

[9] Weiss, B. and Sternlicht, A. (January 8, 2024). *Meta and OpenAI have spawned a wave of AI sex companions—and some of them are children.* Accessed at: https://fortune.com/longform/meta-openai-uncensored-ai-companions-child-pornography/

alone.

7. Definition of “dual-use foundation models”

The EO defines "dual-use foundation model" to mean "an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such as by:

(i) substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons;

(ii) enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber attacks; or

(iii) permitting the evasion of human control or oversight through means of deception or obfuscation."

It should be noted, however, that the broad general purpose capabilities of "foundation" models inherently render them dual-use technologies. These models can often possess latent or unanticipated capabilities, or be used in unanticipated ways that present substantial risk, even if they do not obviously exhibit performance that poses "a serious risk to security, national economic security, national public health or safety, or any combination of those matters" upon initial observation. Furthermore, models that are not developed in accordance with the described characteristics (i.e. trained on broad data, generally using self-supervision, containing at least tens of billions of parameters, and applicable across a wide range of contexts) that exhibit, or can be easily modified to exhibit, high levels of performance at tasks that pose those serious risks should nonetheless be considered dual-use. Novel architectures for AI systems that can be trained on more limited datasets or necessitate fewer parameters, for instance, should fall under the definition if it is evident that they can pose serious risks to national security and public health. Models of this inherently risky architecture AND models that pose an evident risk to security and/or health should be subject to guidance and rigorous safety standards developed by NIST and other agencies pursuant to the EO and beyond.

A slight modification to the EO's definition of "dual-use foundation models," as follows, could accommodate this more inclusive concept of dual-use to appropriately scope NIST's guidance for ensuring the safety of AI systems:

[...]an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; **and** is applicable across a wide range of contexts; ~~*and*~~ **or** that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such…"

8. Global engagement and global military use-cases

Inclusion of strategic competitors: While we welcome efforts through NIST as outlined in Sec. 11 of the EO to advance global technical standards for AI development, we are concerned about the nature of engagement on this issue restricted to 'key international allies and partners'. Cognizant of political realities, we ask that NIST also engage with strategic competitors on global technical standards, in particular those states which are considered to be leaders in AI development, such as the PRC. Without engaging with these strategic competitors, any global standards developed will suffer from a lack of enforcement and global legitimacy. Conversely, standards developed in cooperation with strategic competitors would likely strengthen the legitimacy and enforcement potential of technical standards. Moreover, it is in the United States' national security interests for adversaries' AI to behave more reliably and predictably, and for these systems to remain under proper human control, rather than malfunctioning to escalate situations without human intent or otherwise cause substantial harm that could diffuse beyond their borders.

The exclusion of military AI use-cases will hinder progress on developing global technical standards generally: As the EO outlines, developing global technical standards on civilian AI development and deployment is vital to reaching a global agreement on use of AI. However, considering the blurry boundary between AI developed and deployed for civilian versus military use, we are concerned that a standards agreement on civilian AI alone will likely be difficult without discussing basic guardrails regarding military development and use of AI. This is because with the most advanced AI systems, distinguishing between military and civilian use cases is becoming and will continue to become increasingly difficult, especially considering their general-purpose nature. Mistrust regarding military AI endeavors is likely to impede the international cooperation necessary to ensure global safety in a world with powerful AI systems, including in civilian domains. Adopting basic domestic safety standards for military use of AI, as recommended in #1 ("Military and national security use-cases"), would reduce the risk of catastrophic failure of military systems and inadvertent escalation between strategic competitors, encourage international adoption of military AI safety and security standards, and foster the trust necessary to encourage broader civilian global AI standards adoption. Hence, we reiterate the request that NIST work actively with the Department of State, the Assistant to the President for

National Security and other relevant actors as specified in Section 11, to clarify how its AI safety and security standards can be applied in the military context,, especially with respect to models that meet the EO definition of 'dual-use foundation models'.

CLOSING REMARKS

We appreciate the efforts of NIST to thoughtfully and comprehensively carry out its obligations under the AI EO and are grateful for the opportunity to contribute to this important effort. We hope to continue engaging with this project and subsequent projects seeking to ensure AI does not jeopardize the continued safety, security, and wellbeing of the United States.