FAS FEDERATION OF AMERICAN SCIENTISTS

+1 412 627 7012
DKAUSHIK@FAS.ORG
FAS.ORG

FEDERATION OF AMERICAN SCIENTISTS
1112 16TH STREET NW, SUITE 600
WASHINGTON, DC 20036

July 7, 2023

Dr. Arati Prabhakar
Director
Office of Science and Technology Policy
The White House

RE: Federation of American Scientists' Comment on OSTP-2023-11346; Request for Information; National Priorities for Artificial Intelligence

The Federation of American Scientists (FAS) is a catalytic, non-partisan, and nonprofit organization committed to using science and technology to benefit humanity by delivering on the promise of equitable and impactful policy. FAS believes that society benefits from a federal government that harnesses science, technology, and innovation to meet ambitious policy goals and deliver impact to the public. We are writing today to provide a response to the Office of Science and Technology Policy (OSTP)'s request for information regarding U.S. national priorities and future actions on AI.

Specifically, FAS' comments today will outline specific measures that agencies across the federal government could take to protect the rights and safety of all Americans. As U.S. companies develop powerful and unprecedented frontier AI models, the federal government must take smart steps to steer the development of this technology toward the public good and to mitigate risk. While there are a number of questions this comment could touch upon, we seek to specifically address questions 1, 2, 3, 4, and 7 in the RFI.

Recommendation 1: OSTP should work with a suitable agency to develop and implement a pre-deployment risk assessment protocol that applies to any frontier AI model.

Before the release of a frontier AI system, developers should ensure that the system is sufficiently safe, trustworthy, and reliable. A pre-deployment risk assessment protocol is a potentially powerful tool to evaluate such a system. The goal of this protocol is to rigorously analyze frontier AI models for potential risks, vulnerabilities, and misuse scenarios before deployment. Implementation of such a system would serve as a critical safety practice within our national AI strategy.

<u>Features of a pre-deployment risk assessment protocol</u>

A pre-deployment risk assessment protocol should involve several core features. First, a thorough risk identification process must be established to systematically map out potential threats and vulnerabilities associated with the AI model. Second, the protocol should include a detailed analysis and evaluation of the AI model's capabilities. Finally, the protocol should mandate a robust documentation process for all risk assessment stages. This would include recording all identified risks and the method by which those risks were evaluated, as well as the mitigation measures proposed and their subsequent implementation.

Some such risk assessment tools already exist; for example, the Holistic Evaluation of Language Models from Stanford's Center for Research on Foundation Models "aims to improve the transparency of language models."[1] And these methods are supported by some AI developers who are aware of the need for pre-deployment risk assessment. OpenAI, for example, brought in outside "red-teamers" to find ways that GPT-4 could fail or cause harm, and then used Reinforcement Learning on Human Feedback (RLHF) to "train" bad behaviors out of the model. This strategy, while helpful, was not enough; "the controls put in place are not robust, and methods for mitigating bad model behavior are still leaky and imperfect."[2] This policy could create a more robust and standardized pre-deployment assessment protocol.

Following best practices in other high-risk sectors, these risk assessments could be undertaken by third-party assessment organizations (3PAOs). These bodies, possessing both neutrality and specialized expertise, could enhance the quality and credibility of the risk assessments. For instance, in the FedRAMP program, 3PAOs "perform initial and periodic assessments of cloud systems based on federal security requirements;" these neutral assessors can give an objective accounting of cloud systems' security.[3] Similarly, 3PAOs for AI models' pre-deployment risk assessment would ensure a competent, neutral review.

Applying the NIST AI RMF to pre-deployment risk assessment

The National Institute of Standards and Technology (NIST) AI Risk Management Framework (RMF) includes recommendations and best practices for AI developers to follow to manage and mitigate risk.[4] However, the RMF does not currently provide concrete standards or metrics by which to measure the risk of an AI model before deployment. Our organization has called for more funding for NIST to expand its capacity for risk measurement, and for NIST to develop

[1] Percy, L. et al. (2020). Holistic Evaluation of Language Models. https://arxiv.org/pdf/2211.09110.pdf.
[2] Kaushik, D., & Alexander, L. (11 May 2023). *How Do OpenAI's Efforts to Make GPT-4 'Safer' Stack up against the NIST AI Risk Management Framework?* Federation of American Scientists. fas.org/publication/how-do-openais-efforts-to-make-gpt-4-safer-stack-up-against-the-nist-ai-risk-management-framework/.
[3] *Assessors.* FedRAMP.gov. www.fedramp.gov/assessors/.
[4] (26 January 2023) *AI Risk Management Framework*. National Institute of Standards and Technology. https://www.nist.gov/itl/ai-risk-management-framework.

more concrete benchmarks for assessing an AI model's risk.[5] In developing these benchmarks, NIST would also need to consider various cases, including open-source models, academic research on foundation models, and fine-tuning of AI models, which may not be as clear-cut as the case of large labs' de novo development of frontier AI models such as OpenAI's GPT-4. NIST could also take inspiration from the European Union (EU)'s Assessment List for Trustworthy AI (ALTAI), which was developed alongside the EU AI Act as a voluntary trustworthiness self-assessment tool for AI developers to use prior to deployment of their models.[6]

With a robust set of standards and metrics in place to evaluate the risks of a frontier AI model, the next step is to ensure that this system is consistently implemented for all frontier AI models before deployment. At present, NIST has taken pains to communicate the voluntary nature of the AI RMF, consistent with its Congressional mandate. However, as frontier AI models grow in capabilities, they also grow in risk. Therefore, federal agencies could consider that all frontier AI models developed with federal funding support be subject to a mandated pre-deployment risk assessment protocol.

Defining which AI models and developers to include in this policy would require a multi-dimensional approach. In asserting more concrete metrics and standards for pre-deployment risk assessment, NIST could also outline features of models that should be required to undergo such assessment. For example, labs developing AI with advanced cognitive capabilities in multiple tasks, similar to or surpassing human performance, and employing significant human or computational resources should be included. Additionally, labs whose models are used by a large number of entities, have potential for large-scale misuse, raise privacy or transparency concerns, or primarily focus on AI R&D might qualify for such oversight. The standards for inclusion in the pre-deployment risk assessment policy would need to be flexible to account for compute, algorithmic, and talent efficiency gains over time.

Implementation recommendation: the Federal Trade Commission

[5] Kaushik, D., & Alexander, L. (11 May 2023). *How Do OpenAI's Efforts to Make GPT-4 'Safer' Stack up against the NIST AI Risk Management Framework?* Federation of American Scientists. fas.org/publication/how-do-openais-efforts-to-make-gpt-4-safer-stack-up-against-the-nist-ai-risk-management-framework/.

[6] (17 July 2020) *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment.* Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

We propose that the Federal Trade Commission (FTC) could implement a pre-deployment risk assessment policy. The FTC, under the authority of Section 5 of the FTC Act, has the power to prohibit ”unfair or deceptive acts or practices in or affecting commerce.”[7] If an AI system causes substantial injury to consumers that could not be avoided through reasonable caution, and if the risk of harm is not outweighed by potential benefits to consumers or competition, it might be considered unfair under this authority.

The FTC has issued informal guidance on its “Business Blog” to AI developers. In two posts, the commission reiterates its jurisdiction over commercial AI developers and warns them against unfair or deceptive practices. The FTC has also asserted its enforcement authority over automated systems in a “Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems,” along with the Consumer Financial Protection Bureau, the Equal Employment Opportunity Commission, and the Department of Justice.[8] Since the FTC has warned that discriminatory or biased impact of AI systems is unlawful, it seems natural to extend that interpretation to risky or unsafe AI systems. Hence, the FTC should implement and enforce a pre-deployment risk assessment protocol for frontier AI systems.

Recommendation 2: Adherence to the appropriate risk management framework should be compulsory for any AI-related project that receives federal funding.

The federal government is a very important funder of AI-related projects, through both procurement contracts and grants. According to researchers at Stanford University’s Institute for Human-Centered AI, the U.S. government spent $3.3 billion on AI contracts in fiscal year 2022, largely via the Department of Defense.[9] Additionally, the National Science Foundation (NSF) spends roughly $800 million on AI annually, and $200 million on microelectronics and semiconductors, mostly in the form of grants to academics and research institutions.[10] The NSF also recently established several National AI Research Institutes, which will serve as “hubs for academia, industry and government to accelerate discovery and innovation in AI.”[11] The Department of Energy (DoE) is another important AI-related funder; one key DoE program is the Advanced Scientific Computing Research (ASCR) program, which spends tens of millions of dollars annually to develop “new AI/ML tools that are robust, understandable, and repeatable,”

[7] Federal Trade Commission Act, 15 U.S.C. § 45(a) (1914).
[8] Khan, Lina M. et. al (25 April 2023). *Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems.* Federal Trade Commission. www.ftc.gov/system/files/ftc_gov/pdf/EEOC-CRT-FTC-CFPB-AI-Joint-Statement%28final%29.pdf
[9] Maslej, N., et. al (April 2023). *The AI Index 2023 Annual Report*. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University. https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf.
[10] *FY 2024 Budget.* National Science Foundation. https://new.nsf.gov/budget.
[11] *Artificial Intelligence (AI) at NSF*. National Science Foundation. https://www.nsf.gov/cise/ai.jsp.

among other contributions.[12] Given the large sums of money involved in federal AI contracts and grants, the federal government has both a responsibility to ensure that its AI applications meet a high bar for risk management and an opportunity to enhance a culture of safety in AI development more broadly.

In order to qualify for federal funding, AI projects should be required to adhere to the appropriate risk management framework. Examples of such frameworks include the NIST AI RMF, the DoE's AI Risk Management Playbook[13], and the Defense Intelligence Unit (DIU)'s Responsible AI Guidelines.[14] Currently, these frameworks are only voluntary guidelines and collections of best practices. These frameworks should instead be compulsory, a requirement of any AI-related project seeking federal funding.

For procurement contracts with federal agencies, the contract language should include a requirement for the contractor to comply with the appropriate AI risk management framework. For agencies that do not have their own guidelines, the NIST AI RMF should be used. Agencies should require contractors to document and verify the risk management practices in place for the contract. In the case of grant funding, the NSF should require in grant applications for AI projects documentation of the grantee's compliance with the NIST AI RMF.

Recommendation 3: NSF should increase its funding for "trustworthy AI" R&D

"Trustworthy AI" generally refers to AI systems which are "valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed."[15] Certain research directions can help to promote these properties in AI systems, such as interpretability research, out of domain robustness, fairness and nondiscrimination, and privacy preserving machine learning, among others. By investing substantially in some of these fields, NSF could make useful progress in ensuring that AI systems can be trusted, understood, and monitored.

[12] *FY 2022 Budget Justification.* Department of Energy. https://www.energy.gov/sites/default/files/2022-07/doe-fy2022-budget-vol-2-v3.pdf.
[13] *DOE AI Risk Management Playbook (AIRMP)*. Department of Energy. https://www.energy.gov/ai/doe-ai-risk-management-playbook-airmp.
[14] *Responsible AI Guidelines*. Defense Innovation Unit. https://www.diu.mil/responsible-ai-guidelines.
[15] *Artificial Intelligence*. National Institute of Standards and Technology. https://www.nist.gov/artificial-intelligence.

NSF is the most important non-military funder of AI research and development in the U.S. government,[16] and it plans to spend roughly $800 million on AI in FY24.[17] Per our rough estimates, over the past five years, NSF's investments in these above research areas has stayed relatively flat, making up roughly 10-15% of all AI grant funding from NSF's CISE Directorate. We recommend that NSF shift its strategy and substantially increase the proportion of AI R&D which is directed at some core problems in trustworthy AI in FY24.

To make progress, NSF could take a few complementary steps. NSF could introduce more focused solicitations to advance trustworthy AI, building on existing efforts.[18] Additional programs could focus on some of the research directions mentioned above. NSF could also introduce a new "trustworthy AI" statement in the application process for funding for AI projects, explicitly asking researchers to identify if and how their project contributes to trustworthy AI goals. This statement would allow reviewers to assess advances to AI trustworthiness as part of the "broader impacts" of AI projects. Reviewers would be instructed to favor work which offers a strong case for potential benefits on some of the identified core trustworthy AI research directions. NSF could also encourage or require researchers to follow the NIST AI RMF when conducting their research, as per Recommendation 2.

Recommendation 4: FedRAMP should be broadened to cover AI applications contracted for by the federal government.

The Federal Risk and Authorization Management Program (FedRAMP) is a U.S. government-wide program that standardizes the approach to security assessment, authorization, and continuous monitoring for cloud products and services.[19] The program uses a "do once, use many times" framework to save time and costs associated with conducting redundant agency security assessments. Once a cloud service provider (CSP) is FedRAMP authorized, they have demonstrated adherence to robust security standards, and any federal agency can then adopt their services with confidence in the CSP's security protocols. This streamlines the process for federal agencies to adopt advanced cloud solutions, knowing that the necessary security precautions are in place.

[16] (18 Jan. 2023). *Artificial Intelligence R&D Investments Fiscal Year 2018 - Fiscal Year 2023*. The Networking and Information Technology Research and Development (NITRD) Program. www.nitrd.gov/apps/itdashboard/ai-rd-investments/.
[17] *FY 2024 Budget.* National Science Foundation. https://new.nsf.gov/budget.
[18] A few recent NSF programs can serve as valuable templates: Institute for Trustworthy AI in Law and Society, Safe Learning-Enabled Systems, Fairness in AI.
[19] *Program Basics*. FedRAMP. www.fedramp.gov/program-basics/

Given the rapidly increasing reliance on AI services in federal operations, these services, like federal cloud services, should adhere to robust security standards. Federal agencies manage highly sensitive data related to national security, public services, and individual privacy. The security of AI systems is not only about safeguarding data integrity, but also about ensuring public trust and national safety. Breaches or misuse of these AI services could lead to threats to national security, such as leaks of classified information, violations of privacy rights, or disruptions in essential public services. Thus, stringent security standards are vital to the responsible integration of AI services into federal operations.

Expanding FedRAMP's mandate to include AI services is a logical next step in ensuring the secure integration of advanced technologies into federal operations. Applying a framework like FedRAMP to AI services would involve establishing robust security standards specific to AI, such as secure data handling, model transparency, and robustness against adversarial attacks.

Under this expanded mandate, AI models would undergo a stringent security assessment, authorization, and continuous monitoring process. Like with cloud services, AI models would be assessed by Third Party Assessment Organizations (3PAOs), which would perform security assessments and make recommendations regarding authorization. Once a particular AI model became FedRAMP authorized, its adherence to rigorous security standards would be confirmed. This would streamline the process of integrating AI services into federal operations, saving costs and time by avoiding redundant agency-specific security assessments.

Just as FedRAMP encourages a culture of security in cloud services, its extension to AI services would similarly drive a broader industry trend towards more secure and responsible AI development. Broadening FedRAMP to cover AI would represent a significant stride in the federal government's journey towards secure and responsible AI utilization.

Recommendation 5: The Department of Homeland Security should establish an AI incidents database.

The Department of Homeland Security (DHS) should track AI-related harms in a centralized AI Incidents Database. This database would serve as a central repository for reported AI-related incidents across industries and sectors in the United States, ensuring that crucial information on breaches, system failures, misuse, and unexpected behavior of AI systems is systematically collected, categorized, and made accessible.

The DHS is well-positioned to maintain such a database, given its mission to ensure the safety and security of the country. Under the Homeland Security Act of 2002 (Pub. L. No. 107-296), the DHS has broad authorization to perform actions necessary to protect the safety and security of the United States, including prevention, preparedness, response, and recovery from both natural disasters and man-made threats. This authorization includes the ability to collect, retain, analyze, and disseminate information relevant to these threats.

Specifically, Section 201(d) of the Act charges the Secretary of the Department of Homeland Security with responsibilities that include accessing, receiving, and analyzing law enforcement information, intelligence information, and other information from agencies of the federal government, state and local government agencies, and private sector entities, and integrating that information to protect against terrorism.[20] The Cybersecurity and Infrastructure Security Agency (CISA) has a current mandate established in § 2209 of the Homeland Security Act ”to provide analysis on cyber threat information…designated by the Secretary.” Building upon these foundations, AI-related reporting requirements can be incorporated into existing guidelines like the National Cyber Incident Response Plan (NCIRP). Moreover, sector-wide voluntary initiatives such as the Financial Services Information Sharing and Analysis Center (FS-ISAC) can also serve as templates to adapt incident reporting processes for AI applications.

The database should be designed to encourage voluntary reporting from AI developers, operators, and users while ensuring the confidentiality of sensitive information. Furthermore, the database should include a mechanism for sharing anonymized or aggregated data with AI developers, researchers, and policymakers to help them better understand and mitigate AI-related risks. The DHS could build on the efforts of other privately collected databases of AI incidents, including the AI Incident Database created by the Partnership on AI and the Center for Security and Emerging Technologies.[21] This database could also take inspiration from other incident databases maintained by federal agencies, including the National Transportation Safety Board’s database on aviation accidents.[22]

The DHS should collaborate with NIST in designing and maintaining the database, including setting up protocols for data validation, categorization, anonymization, and dissemination. Additionally, it should work closely with AI industry stakeholders, academia, and civil society to ensure that the database is comprehensive, useful, and trusted by stakeholders.

[20] Homeland Security Act, 6 U.S.C. § 121(d) (2002).

[21] *The First Taxonomy of AI Incidents*. AI Incident Database. incidentdatabase.ai/blog/the-first-taxonomy-of-ai-incidents/.

[22] *Aviation Accidents - Index of Months*. National Transportation Safety Board. www.ntsb.gov/Pages/monthly.aspx.

By providing a clear and comprehensive view of AI-related incidents, this database will enhance the collective understanding of the safety and security landscape surrounding AI systems.

Recommendation 6: OSTP should work with agencies to streamline the process of granting Interested Agency Waivers to AI researchers on J-1 visas.

The National Security Commission on Artificial Intelligence emphasized the central role talent will play in the ongoing rivalry surrounding artificial intelligence technology.[23] They stated, "The winner of the AI competition will not be determined solely by superior technology but also by the side with access to a diverse and highly skilled pool of tech-savvy talents." The United States is currently engaged in an intense global contest to attract and maintain scarce AI experts.

The J-1 Exchange Visitor Program is a visa pathway that allows visitors to come to the United States and is often utilized by visiting researchers and postdocs, yet it requires some of them to return home for two years before they can apply for a different status that permits permanent residency and work. This requirement applies in cases where the visitor's program skills are recognized as necessary for their home country's economic growth or if the visit was funded by either the US or the home country.

The administration clearly sees the J-1 program as an effective way to attract STEM talents, particularly through the Early Career STEM Initiative. However, the two-year home residency requirement restricts the seamless transition to permanent residency, preventing skilled J-1 beneficiaries from furthering scientific and technological development in America.

In some circumstances, federal agencies are authorized to request a waiver of this two-year requirement when they have interest in particular J-1 visa holders via an "Interested Government Agency" (IGA) request.

For candidates involved in AI research, there should be a process established for applying for relevant agencies to act as IGAs and arrange waivers. The current method lacks transparency and predictability; introducing a structured system with published eligibility criteria could significantly enhance the process. OSTP should work with agencies to establish a streamlined process. Agencies can take inspiration from the Department of Defense, which maintains a

[23] *(2021) Final Report*. National Security Commission on Artificial Intelligence. https://www.nscai.gov/2021-final-report/

dedicated webpage outlining their application process, providing useful resources like an application checklist and sample sponsor letter.[24]

Conclusion

Artificial intelligence presents great opportunities and potential challenges. In order to harness this technology for the public good, agencies across the federal government must both promote the trustworthy development and use of AI and mitigate its risks.

The actions we have recommended would all help build American capacity to encourage development and use of AI systems for the public good while mitigating risks. These actions would be timely; in most cases, they could be implemented by the relevant agencies under existing authorities. They would help align the incentives of AI developers with the public good by tying funding and contracts to stringent security protocols and proper risk management. And they would help inform policymakers and the public of AI-related incidents.

We thank you for considering our insight on these issues and our recommendations for the National AI Strategy. If you have any questions, please feel free to reach out at dkaushik@fas.org.

Sincerely,

Divyansh Kaushik
Associate Director, Emerging Technologies and National Security
Federation of American Scientists

Jack Cunningham
Fellow, Federation of American Scientists

Liam Alexander
Fellow, Federation of American Scientists

[24] *USD (R&E), "DoD J1 Visa Waiver Program."*
https://basicresearch.defense.gov/Programs/DoD-J1-Visa-Waiver-Program/