# PUBLIC SUBMISSION

**Received:** May 28, 2025
**Tracking No.** mb8-gg4e-bc2n
**Comments Due:** May 28, 2025 **Submission Type:** API

**Docket:** NSF-2025-OGC-0001
NITRD_FRDOC_0001

**Comment On:** NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

**Document:** NSF-2025-OGC-0001-DRAFT-0162
Comment on FR Doc # 2025-07332

## Submitter Information

**Organization:** Center for Security and Emerging Technology (CSET), Georgetown University

## General Comment

The comment from the Center for Security and Emerging Technology is attached.

## Attachments

CSET_RFI_Response

**RFI Response: The Development of a 2025 National Artificial Intelligence (AI) Research and Development (R&D) Strategic Plan**

**Federal Register Document Citation**: 90 FR 17835
**Agency Name**: Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO), National Science Foundation
**Organization**: Center for Security and Emerging Technology (CSET), Georgetown University
**Primary Point of Contact:** Kendrea Beers

The Center for Security and Emerging Technology (CSET) at Georgetown University offers the following comments in response to the Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO) and National Science Foundation's request for information on **the Development of a 2025 National Artificial Intelligence (AI) Research and Development (R&D) Strategic Plan.**

A policy research organization within Georgetown University's Walsh School of Foreign Service, CSET provides decision-makers with data-driven analysis on the security implications of emerging technologies, focusing on artificial intelligence, advanced computing, and biotechnology. We appreciate the opportunity to offer these comments.

This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in developing the 2025 National AI R&D Strategic Plan and associated documents without attribution.

# Overview

Georgetown University's Center for Security and Emerging Technology (CSET) presents the following recommendations on the development of a 2025 National Artificial Intelligence (AI) Research and Development (R&D) Strategic Plan. These recommendations are drawn from CSET's extensive research portfolio. For reference and ease of use, our recommendations correspond to specific strategies enumerated in the May 2023 update to the National AI R&D Strategic Plan, whose rewriting this RFI seeks to inform.

We primarily focus our recommendations on rewriting aspects of four selected strategies from the 2023 update. For Strategy 1, "Make Long-Term Investments in Fundamental and Responsible AI Research," we recommend emphasizing the urgent need for the U.S. government to fund research into hardware optimized for AI algorithms. For Strategy 4, "Ensure the Security of AI Systems," we recommend sharpening focus on research to mitigate risks from frontier AI systems. For Strategy 6, "Measure and Evaluate AI Systems through Standards and Benchmarks," we recommend promoting a science of measuring AI rather than focusing too narrowly on existing measurement techniques. For Strategy 7, "Better Understand the National AI R&D Workforce Needs," we make targeted recommendations for AI workforce priorities guided by our data-driven research. In addition, we recommend that a new AI R&D strategy leverage open source analysis to eliminate information gaps around cutting-edge AI developments for policymakers and also to benefit research security.

# Strategy 1: Make Long-Term Investments in Fundamental and Responsible AI Research

## *Maintain American leadership in AI hardware*

- **Fund research into hardware optimized for AI.** The section "Advancing Hardware for Improved AI" should stress the need for the United States government to both build on American strengths in current hardware paradigms and explore other paradigms. For example, the 2023 update cites neuromorphic processors as one example of hardware optimized for AI algorithms; CSET data shows that neuromorphic processor research is one of the largest and fastest-growing sectors of chip design and fabrication research and is currently dominated by China. Compared to the 2023 update, the 2025 strategy should

emphasize the urgent need for the U.S. government to fund research into hardware optimized for AI algorithms (including but not limited to neuromorphic processors) to preserve American leadership in AI hardware.

# Strategy 4: Ensure the Security of AI Systems

The 2023 update emphasizes risks and security challenges that are most relevant to classical machine learning techniques developed decades ago. While research in these areas is still valuable, the most pressing risk mitigation and security challenges stem from the most advanced general-purpose AI systems, i.e. frontier models. The 2025 strategy should emphasize security research to mitigate the greatest risks to national security from frontier AI.

## *Re-focus on AI risk mitigation research priorities*

This section should be updated to align with expert consensus on AI security and risk mitigation research priorities, which fall into the following three high-level categories.

- **Support a science of assessing risk.** The 2023 update is correct to emphasize risk assessment as a core priority. Since 2023, frontier AI companies have developed frameworks for mitigating catastrophic risks that rely heavily on capability evaluations in areas such as cyber, biosecurity, and AI R&D. For these frameworks to be effective, the evaluations need to be rigorous, as discussed further in our recommendations for Strategy 6. Research is also needed to improve the rigor of broader risk assessment, including investigating lesser-known threat models such as risks from internal deployment of AI systems.

- **Improve methods at the development stage to build trustworthy AI systems.** Since 2023, AI misalignment has shifted from theoretical concern to empirical reality: researchers have documented AI models engaging in reward hacking, scheming, alignment faking, and other misbehaviors. Research in areas including robustness, interpretability, and truthfulness is urgently needed to understand and mitigate these behaviors.

- **Develop methods at the deployment stage to prevent AI systems from causing harm even when they behave more dangerously in deployment than in testing.** Systems' behavior can be more dangerous in deployment for several well-documented reasons: jailbreaking by users, robustness failures, alignment faking, or data poisoning by malicious actors to create "sleeper agents." The federal government should prioritize funding the nascent field of AI control, which aims to mitigate these risks during deployment via techniques drawn from cybersecurity and software engineering as well as novel techniques that leverage AI systems to provide oversight.

## *Prioritize securing frontier AI systems*

Under the heading "Securing AI," securing frontier AI systems should be the top priority.

- **Secure frontier model weights.** Most importantly, the federal government needs to coordinate a targeted R&D effort to develop the option to secure frontier model weights against theft by state-level actors. This includes discrete lines of technical research that can be advanced in parallel, such as scaling confidential computing and developing open-source secure API designs. Research to secure frontier AI models should also address the threat model of self-exfiltration or rogue internal deployment by advanced AI systems that are compromised (e.g., sabotaged by adversaries or misaligned).

- **Understand threats to frontier AI development.** Beyond securing frontier models specifically, the 2025 strategy should also recommend research to explore the possibility of sabotage of frontier AI development in general. For example, it is important to understand how feasible it would be for adversaries or compromised AI systems to subtly degrade codebases or insert backdoors into next-generation frontier models.

- **Understand threats to AI components in critical infrastructure.** Related research should study the prospects of securing AI systems that are integrated into critical infrastructure, such as AI systems that are themselves used for cyberdefense.

## *Harness AI for cybersecurity*

The 2025 strategy should include a new top-level section as a key funding priority: "Harnessing AI for Cybersecurity." AI has the potential to significantly advance cyber operations: American companies are using general-purpose AI systems to automate key cybersecurity tasks; threat actors are leveraging frontier AI for large-scale cyber operations, especially via social engineering; and machine learning systems are already important components of cyberdefense workflows for tasks like anomaly detection. However, significant work is needed to realize the potential of AI for cybersecurity.

- **Support a variety of approaches to AI for cybersecurity.** Federal research funding should support evaluating current AI cyber capabilities, developing AI tools for cyber defenders, analyzing the reliability of these tools, using AI to generate formally verified code, and beyond.

# Strategy 6: Measure and Evaluate AI Systems through Standards and Benchmarks

The 2025 strategy should better reflect the need to fund a broader science of measuring AI, not limited by existing tools under the categories of standards and benchmarks.

## *Clarify terminology around AI measurement*

- **Clarify the relationship between evaluations, benchmarks, and other tools for measurement.** Compared to the 2023 update, the 2025 strategy should distinguish more clearly between terms. For instance, the definition of "benchmark" in the 2023 document is too narrow. We recommend that the new strategy define a benchmark as "a standardized AI evaluation that measures system performance on specific tasks using consistent datasets, metrics, and protocols."

  In addition, metrics and benchmarks are not always appropriate nor sufficient for evaluating AI system capabilities and effectiveness. Other evaluations such as red-teaming and operational evaluations can provide a more thorough picture of AI system performance and also deserve investment and attention.

- **Refine recommendations for testing requirements.** We recommend incorporating language into the strategy that refines the current discussion of testing requirements. E.g., "Testing requirements should ensure that trustworthy AI systems accord with designer specifications and also achieve their intended outcomes in the real world."

- **Clarify certification needs in the new strategy.** Certifications of AI ethical assurance can help build confidence in AI systems. However, certifications must be conducted in consistent ways and performed by trusted authorities in order to be effective. Research on standardizing certifications for AI could take inspiration from other domains, such as the Leadership in Energy and Environmental Design program for certifying buildings' sustainability.

## *Allow for definitional flexibility across differing application areas*

- **Acknowledge domain differences in AI definitions.** The 2023 update suggests that all AI practitioners must adhere to the same technical definitions for AI terms, which seems infeasible given the many different contexts in which AI is used. We recommend including language in the new strategy that acknowledges the need for some flexibility in definitions. E.g., "There is a need to achieve consensus-based definitions of technical terms and consistent terminology (e.g., AI, autonomy, transparency, explainability, and interpretability) within certain domains and application areas while acknowledging that users in these areas may define terms differently."

### *Collect AI incident data to inform policy*

- **Support incident reporting.** Standardized reporting of AI accidents and security incidents, or events where AI systems cause harm or are subverted in some way, provides decision-makers with important data and evidence when determining which harms are most pressing, whether existing mitigations or interventions are effective, and whether emergent risks demand immediate attention. However, schemas for incident reporting are currently diffuse and inconsistent, making it difficult for groups who have adopted different schemas to share insights on AI risks and harms. The 2025 strategy should recommend research to help standardize components of incident reports and establish a common baseline of fields that should be collected in every incident report.

## Strategy 7: Better Understand the National AI R&D Workforce Needs

CSET's data-driven research points toward improvements to this section: clarifying the role of foreign AI workers, supporting education for AI ethics and governance occupations, and leveraging lesser-known pathways for AI skill-building.

### *Acknowledge contributions of foreign workers to AI research*

- **Attract and retain both foreign and domestic talent.** The 2023 Strategy, under the heading, "Identifying and Attracting the World's Best Talent," mentions declining graduate enrollment among permanent residents and citizens in AI related programs. It goes on to state that half of AI experts in industry and academia are born outside of the United States. The 2025 strategy should explicitly recognize that maintaining R&D leadership requires foreign students and workers with skills in AI. Rather than mentioning existing federal resources and international partnerships with universities and governments, the strategy should proactively focus on attracting and retaining foreign students and workers.

  Beyond temporary visas for work and study, AI experts need specific pathways to citizenship so that they can take jobs that require citizenship or permanent residency. Higher education reforms should be made to attract more domestic graduate students to AI degree programs, but these reforms alone will not be able to displace the contributions of foreign workers to AI research in the United States.

### *Make data-driven recommendations for AI education*

- **Support education for AI ethics and governance occupations through a data-driven approach.** The 2023 report section titled "Incorporating Ethical, Legal, and Societal Implications into AI Education and Training" recommends that the federal government support educational programs designed to build interdisciplinary competencies, and support

dissemination of education materials on ethical, legal and social aspects of AI for integration in AI education and training programs.

To expand on this focus in the 2025 strategy and achieve these goals, CSET recommends using real-time labor market information to identify in-demand skills for AI ethics and AI governance occupations. Our analysis of job postings data found that the skill compositions for ethics and governance occupations differ, although both are multi-disciplinary. While the most prominent skills for AI ethics occupations are technical expertise, research skills and policy analysis, AI governance occupations require project management, business leadership and operations, and legal compliance, and demand for these occupations varies across sectors. Incorporating a data-driven approach into the strategy can help institutions of higher education design targeted and multifaceted curricula and enable government agencies to better align with the motivations of individual learners and employers.

- **Take advantage of registered apprenticeships and community colleges for AI education.** We recommend adding the following recommendation to the 2025 strategy in any sections that correspond to the 2023 update's "Training/Retraining the Workforce" heading:

  "Registered apprenticeships in AI-related occupations and AI-related programs at community colleges are two pathways for training and retraining workers that align with these objectives. Registrations of new apprentices in AI-related occupations have steadily increased over the past decade, and continued support from Congress and the administration is needed to continue that trend. Community colleges often serve as providers of related technical instruction for apprenticeship programs, and in partnership with industry a growing number of community colleges have developed their own AI coursework. Incentivizing partnerships with employers and providing increased funding for community colleges in general would help expand this training pipeline."

# Leveraging Open Source Analysis to Benefit AI R&D

A key aspect of the AI R&D strategy not addressed in the 2023 update involves leveraging open source analysis to enhance government awareness of state-of-the-art developments in research at home and abroad. In a rewritten 2025 AI R&D strategy, such analysis should play a key role in identifying and analyzing cutting-edge research and promoting research security. Since AI is being developed, deployed, and used almost entirely outside of the federal government, an open source capability could provide reliable information about the current state and future of AI technology and help alleviate information disadvantages for policymakers and researchers.

## *Use open source to enhance research security and horizon-scanning*

- **Significantly expand open-source intelligence (OSINT) gathering and analysis on AI.** There is currently no office or agency, inside or outside government, that can provide a comprehensive view of the AI landscape, and the intelligence community remains squarely focused on classified sources. OSINT collection related to basic research is critically underdeveloped and under-resourced in the federal government. Significant investments are needed in collection, interpretation, and dissemination of AI OSINT, incorporating sources like research publications and workforce data. Such intelligence may be used to identify promising AI trends that deserve further exploration or investment or that may carry implications for national security.

  For instance, such a capability would enhance the ability to understand how developments in competitor states, like China, might affect the U.S. R&D landscape. The lack of a serious program to track China's AI progress undermines federal efforts across policy domains, including research security and industrial policy, and raises the risk of technological surprise. An enhanced open source capability would play an important role in monitoring and understanding China's AI ecosystem, including the role of the Chinese government itself, related actors such as state-owned enterprises, state research labs, and state-sponsored technology investment funds, and other actors, such as universities and tech companies, with whom American researchers might collaborate.

## *Exploit information sharing with the private sector and allies*

- **Establish reporting programs to gather information on AI development processes from AI companies**. Reporting programs could ask companies to provide detailed documentation on training procedures and environments, results of capability and risk evaluations on frontier models, model specifications (also known as constitutions) that define the behaviors that companies want AI models to have, and explanations of why AI companies believe their current risk management practices suffice. Detailed documentation on AI development practices would decrease the information gap between AI developers and the government, enabling policy to quickly respond to increasing AI capabilities. For this particular information, the public or governmental interest in transparency greatly exceeds the potential downside (for national security or for companies' intellectual property) of making the information available to competitors.

- **Contribute to and draw from the collective intelligence of U.S. allies regarding AI capabilities.** The impacts of AI systems transcend national borders, and advances in AI R&D in allied countries may also be relevant at home. The U.S. government should draw on information about frontier AI capabilities from trusted allies like the United Kingdom and its AI Security Institute and share information with them to maintain trust.