# PUBLIC SUBMISSION

**Received:** May 29, 2025
**Tracking No.** mb9-qcw7-rtbk
**Comments Due:** May 28, 2025 **Submission Type:** Web

**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-0228
Comment on FR Doc # 2025-07332

## Submitter Information

**Organization:** Center for a New American Security

## General Comment

See attached file for the Center for a New American Security's comment.

## Attachments

CNAS AI RD Strategic Plan Comments

Center for a New American Security

# Securing America's AI Future: Federal Research and Development Priorities

May 29, 2025

The Center for a New American Security (CNAS) welcomes the opportunity to provide a response to the White House Office of Science and Technology Policy's Request for Information (RFI) regarding the development of a strategic plan for artificial intelligence (AI) research and development (R&D). 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.

This submission reflects the views of the following authors:

- Caleb Withers, Research Associate, Technology and National Security Program
- Spencer Michaels, Joseph S. Nye, Jr. Intern, Technology and National Security Program

With thanks to Paul Scharre, Vivek Chilukuri and Janet Egan for their valuable feedback.

## Introduction

Cutting-edge R&D is a pillar of America's AI leadership. Private capital is essential to driving AI R&D progress, with U.S. firms investing an estimated $109.1 billion in 2024—more than twelve times the estimated $9.3 billion from Chinese firms that same year.[1]

Although private capital is essential to maintaining America's edge in AI R&D, it is insufficient. Certain categories of research will remain underfunded when guided solely by commercial incentives. For example, private firms are reluctant to invest in areas with uncertain returns or insufficient value for investors to capture. In these areas, targeted federal investment plays a vital and complementary role. As the Trump administration develops its 2025 National AI R&D Strategic Plan, it should prioritize efforts that:

- Inform U.S. policy development—for example, independent model evaluations to avoid "self-grading" from private AI labs, or independent research into trends in hardware and algorithmic progress, data bottlenecks, distributed training, and model distillation, insights from which can help calibrate relevant policies around export controls;

[1] *Artificial Intelligence Index Report 2025* (Stanford Institute for Human-Centered AI, 2025), https://hai-production.s3.amazonaws.com/files/hai_ai_index_report_2025.pdf, 17.

- Strengthen U.S. policy implementation—for example, research into hardware-enabled governance mechanisms in AI chips to strengthen export controls and associated defenses against sophisticated adversaries with physical access;[2]
- Lack commercial incentives for prioritization or publication, such as surfacing frontier model weaknesses, limitations, and risks, or exploring interactions between models from different companies;
- Support foundational AI research in areas such as interpretability, which may enable future breakthroughs but lack near-term commercial payoff.[3]

The specific recommendations that follow identify areas where federal agencies should expand, initiate, or sustain investment to strengthen U.S. leadership in AI R&D.

**Robustness and Reliability for National Security**

Federal agencies should fund **high-risk, high-reward projects to drive fundamental advances in AI robustness, reliability, interpretability, and explainability**. Frontier AI systems offer enormous potential for analyzing data at speed and scale across cyber, military, and intelligence operations. But significant challenges remain in achieving the robustness and reliability required for transformative national security applications, especially as adversaries seek to compromise or exploit these systems.

The lack of interpretability also creates significant operational risks. Military commanders and intelligence analysts must be able to understand why AI systems recommend certain courses of action, particularly in high-stakes scenarios where lives and national interests are at stake. Labs may underinvest in certain techniques to the depth required for national security applications.[4] Federal R&D is essential to develop systems with the robustness and reliability required for national security missions. Challenges of interpretability and explainability also hinder broad-based adoption in the public sector.

In partnership with leading labs, federal investment should prioritize research addressing:

- Advanced techniques for adversarial testing and stress-testing AI systems to identify vulnerabilities before deployment in critical applications;
- Technical approaches to systematically detect and reduce hallucinations in large language models, including methods that can verify outputs against reliable knowledge bases;
- Methodologies for ensuring consistent performance across diverse deployment contexts, particularly under stress conditions or when processing unexpected inputs;

[2] Tim Fist, Tao Burga, and Vivek Chilukuri, *Technology to Secure the AI Chip Supply Chain: A Working Paper* (CNAS, December 11, 2024), https://www.cnas.org/publications/reports/technology-to-secure-the-ai-chip-supply-chain-a-primer.
[3] Dario Amodei, "The Urgency of Interpretability," darioamodei.com, April 2025, https://www.darioamodei.com/post/the-urgency-of-interpretability.
[4] Dan Hendrycks and Laura Hiscott, "The Misguided Quest for Mechanistic AI Interpretability," AI Frontiers, May 15, 2025, https://www.ai-frontiers.org/articles/the-misguided-quest-for-mechanistic-ai-interpretability.

- Development of comprehensive reliability metrics that capture real-world performance requirements rather than merely benchmark performance;
- Interpretability and explainability to enable trust in, and accountability for, model outputs and to identify unintended or unexpected behaviors.

**AI Standards and Benchmarks**

As AI systems advance and proliferate, establishing standards and benchmarks becomes critical for evaluating capabilities, ensuring interoperability, and building trust. Countries that lead in technical and policy frameworks will have stronger influence over international standards, conferring significant economic and political benefits. The United States must **lead in developing comprehensive evaluation frameworks** for highly capable AI systems that keep pace with technological progress.

Current evaluation approaches face a fundamental challenge: many tests quickly become obsolete.[5] Benchmarks that once differentiated frontier models become saturated within months, eliminating meaningful performance comparisons.[6] In June 2024, Anthropic's Sonnet 3.5 scored 32% on the SWE-bench Verified benchmark; its successor, Sonnet 4, released in May 2025, scored 61%.[7] This rapid capability growth requires evaluation frameworks that adapt as models improve.

Beyond the challenge of obsolescence, existing evaluations often lack the contextual grounding needed for actionable assessments. Effective evaluation must ground abstract performance metrics in operational realities—measuring not just what AI systems can do, but how those capabilities translate into genuine advantages or risks compared to existing alternatives. Additionally, understanding capability *trends*—and the extent they appear sustainable—is arguably as important as measuring current performance.

Inconsistency across industry compounds these problems. While model cards document AI systems, companies use different benchmarks and risk assessment methodologies, making meaningful comparisons difficult. Each leading AI lab employs its own approach to measuring capabilities and risks.[8] Without common evaluation methods, policymakers cannot confidently and efficiently assess AI systems or make informed decisions about deployment in sensitive applications.

---

[5] Anka Reuel et al., "BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices," *arXiv*, 20 November, 2024, https://arxiv.org/html/2411.12990v1.
[6] Tiernan Ray, "AI Isn't Hitting a Wall, It's Just Getting Too Smart for Benchmarks, Says Anthropic," ZDNet, November 22, 2024, https://www.zdnet.com/article/ai-isnt-hitting-a-wall-its-just-getting-too-smart-for-benchmarks-says-anthropic/.
[7] "Introducing Computer Use, a New Claude 3.5 Sonnet, and Claude 3.5 Haiku," Anthropic, October 22, 2024, https://www.anthropic.com/news/3-5-models-and-computer-use; and "System Card: Claude Opus 4 & Claude Sonnet 4," Anthropic, May 2025, https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf.
[8] "xAI Risk Management Framework (Draft)," xAI, February 20, 2025, https://x.ai/documents/2025.02.20-RMF-Draft.pdf; Anca Dragan, Helen King and Allan Dafoe, "Frontier Safety Framework," Google DeepMind, May 17 2024, https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/introducing-the-frontier-safety-framework/fsf-technical-report.pdf; "Preparedness Framework (Beta)," OpenAI, December 18, 2023, https://cdn.openai.com/openai-preparedness-framework-beta.pdf; "Responsible Scaling Policy," Anthropic, March 31, 2025, https://www-cdn.anthropic.com/17310f6d70ae5627f55313ed067afc1a762a4068.pdf.

These measurement gaps have strategic implications. Without standardized evaluation frameworks, the United States will struggle to effectively monitor the global AI capability landscape, assess emerging dual-use risks, or coordinate responses with allies. Private sector evaluation efforts, while valuable, may not adequately address national security considerations or ensure coverage of capabilities relevant to critical infrastructure and defense applications. The sensitive nature of many government and defense applications demands sophisticated approaches to evaluation that can protect classified information and proprietary technologies while still enabling rigorous assessment. This is a technical challenge that merits further research.[9]

Leading developers have little incentive to publicize evaluations that surface their systems' flaws or limit commercial applications. Standardized evaluations also require coordination across firms that otherwise compete on proprietary performance metrics. Only federal R&D can provide the institutional neutrality, continuity, and classification access needed for these efforts. Federal investment should focus on developing evaluation science: the foundational research needed to create robust, adaptive measurement frameworks. This includes research into dynamic benchmarking methodologies that can evolve with advancing capabilities and technical standards that enable consistent evaluation across different AI systems and deployment contexts. Such foundational work requires long-term investment with uncertain commercial payoffs, making it well-suited for federal R&D priorities.

**Support and Expand the National Artificial Intelligence Research Resource (NAIRR)**

As the lead federal agency for computer and information science research, the U.S. National Science Foundation (NSF) should support the AI research community. The NSF can help by providing AI researchers with the infrastructure necessary to conduct cutting-edge AI R&D. **NSF should maintain Strategy 5 of the 2023 Strategic Plan.** By providing research infrastructure to universities, smaller companies, and independent researchers, federal investment can preserve the competitive innovation ecosystem that has long been America's strategic advantage.

Cutting-edge AI R&D often requires large amounts of expensive computing hardware. Hourly rental costs per H100 GPU on leading providers can range between $2 and $12.[10] Training Gemini 1.0 Ultra on cloud computing would have cost an estimated $191 million.[11] These costs create significant barriers for researchers.

---

[9] Andrew Trask et al., "Secure Enclaves for AI Evaluation," OpenMined, November 20, 2024, https://openmined.org/blog/secure-enclaves-for-ai-evaluation/; Lennart Heim, "A Trusted AI Compute Cluster for AI Verification and Evaluation," March 31, 2024, https://blog.heim.xyz/a-trusted-ai-compute-cluster/; Benjamin Bucknall and Robert Trager, *Structured Access for Third-Party Research on Frontier AI Models: Investigating Researchers' Model Access Requirements* (Centre for the Governance of AI, October 2023), https://cdn.governance.ai/Structured_Access_for_Third-Party_Research.pdf; and Yoshua Bengio et al., *International AI Safety Report* (UK Department for Science, Innovation and Technology, January 2025), https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf, sec. 3.4.3.

[10] "NVIDIA H100 Pricing (May 2025): Cheapest On-Demand Cloud GPU Rates," Thunder Compute, May 19, 2025, https://www.thundercompute.com/blog/nvidia-h100-pricing.

[11] Ben Cottier et al., "The Rising Costs of Training Frontier AI Models," *arXiv*, May 31, 2024, https://arxiv.org/abs/2405.21015.

NSF should continue to support and expand the National Artificial Intelligence Research Resource (NAIRR) Pilot in the Strategic Plan.[12] While the compute resources available through NAIRR represent only a fraction of what leading industry players can access, this initiative serves as a crucial first step in democratizing compute access for research institutions across America.[13]

To maintain America's AI leadership, NSF should expand NAIRR's compute capacity and access mechanisms. Under NAIRR, NSF should also develop additional infrastructure support mechanisms such as grant programs, dedicated funding streams, and institutional incentives to bridge the gap between academic and industrial AI research capabilities.

**Anticipating Rapid AI Progress**

The federal government should invest in research that anticipates continued rapid progress in AI capabilities. Recent years have demonstrated a consistent pattern: AI models have repeatedly performed tasks once predicted as impossible. Moreover, while costs to train and deploy the most capable AI models continue to rise, prices for AI capabilities rapidly fall once they have been demonstrated.[14] Although this pace of progress is not guaranteed going forward, **research investments that anticipate continued progress are particularly valuable because they deliver the highest impacts in precisely the scenarios where AI's impacts prove most transformative**. For example, federal AI research might:

- Focus on evaluating capabilities where current models have only nascent abilities—such as end-to-end cyber operations or complex scientific reasoning. Research approaches might include staged assessments, elicitation techniques, and benchmarks that establish human baselines even when current models score near zero;
- Experiment with AI in the research process so systems and researchers are ready to scale their use as capabilities advance, such as AI-driven research agents, self-improving workflows, and methods of maintaining epistemic rigor;[15]
- Advance techniques to increase the ability of government and private sector partners to understand emerging capabilities;[16]
- Prioritize R&D into information security architectures tailored to sensitive AI development and deployments—such as privacy-preserving computing, and broader AI-tailored

[12] "National Artificial Intelligence Research Resource (NAIRR) Pilot," NAIRR, https://nairrpilot.org/.
[13] Kyle Miller and Rebecca Gelles, "The NAIRR Pilot: Estimating Compute," Center for Security and Emerging Technology (CSET), May 8, 2024, https://cset.georgetown.edu/article/the-nairr-pilot-estimating-compute/.
[14] *Artificial Intelligence Index Report 2025*.
[15] Hanchen Wang et al., "Scientific Discovery in the Age of Artificial Intelligence," *Nature* 620 (August 2, 2023): 47–60, https://www.nature.com/articles/s41586-023-06221-2.
[16] See for example "How to Evaluate Control Measures for AI Agents?," United Kingdom AI Security Institute, April 11, 2025, https://www.aisi.gov.uk/work/how-to-evaluate-control-measures-for-ai-agents.

information security techniques that may not be justified by the sensitivity of current AI systems;[17]

- Focus on large, complex models that demonstrate the most advanced capabilities, even as smaller models provide clearer experimental results.[18]

**AI-Biotechnology Nexus**

Harnessing the convergence of AI and biotechnology is both a transformative opportunity and national security imperative. AI has demonstrated remarkable success in accelerating drug discovery, predicting protein structures, identifying novel drug targets, and optimizing clinical trial design—capabilities with both commercial and strategic significance.[19] AI-enhanced biomanufacturing processes promise to revolutionize production of critical goods, from pharmaceuticals to sustainable fuels. To harness AI's potential to accelerate biotechnology breakthroughs, the government must **expand investment in AI-biotechnology research and infrastructure**.

Private-sector biotech firms focus on profit-maximizing markets and may underinvest in low-margin or high-risk R&D. Federally funded AI R&D can fill these gaps, especially in developing applications in the public interest, such as biothreat detection, biosafety systems, and rare disease modeling. In a geopolitical crisis, these are precisely the capabilities the market is unlikely to deliver in time. The People's Republic of China has pursued biotechnology as a strategic priority for two decades through massive state investments and aggressive intellectual property acquisition.[20] The National Security Commission on Emerging Biotechnology warns that China is "dangerously close" to overtaking U.S. biotechnology leadership.[21]

---

[17] Bengio et al., *International AI Safety Report*, sec. 3.4.3.; Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation* (arXiv, February 2018), 96, https://arxiv.org/abs/1802.07228; and *Fast-Track Action Committee on Advancing Privacy-Preserving Data Sharing and Analytics, Networking and Information Technology Research and Development Subcommittee of the National Science and Technology Council, National Strategy to Advance Privacy-Preserving Data Sharing and Analytics* (National Coordination Office, Networking and Information Technology Research and Development Program, March 2023), https://www.nitrd.gov/pubs/National-Strategy-to-Advance-Privacy-Preserving-Data-Sharing-and-Analytics.pdf; Sella Nevo et al., "A Playbook for Securing AI Model Weights," RAND, May 30, 2024, https://www.rand.org/pubs/research_briefs/RBA2849-1.html; and Robert Podschwadt et al., "A Survey of Deep Learning Architectures for Privacy-Preserving Machine Learning With Fully Homomorphic Encryption," *IEEE Access* 10, no. 1 (November 3, 2022), https://ieeexplore.ieee.org/document/9936637.

[18] See, for example Hendrycks and Hiscott, "The Misguided Quest for Mechanistic AI Interpretability."

[19] "Exploring the AI impact on biotech advancements," WHX Insights, May 26, 2024, https://www.worldhealthexpo.com/insights/ai-automation/exploring-the-ai-impact-on-biotech-advancements; John Jumper et al., "Highly Accurate Protein Structure Prediction with AlphaFold," *Nature* 596 (July 15, 2021): 583–589, https://www.nature.com/articles/s41586-021-03819-2.; Frank Pun, Ivan Ozerov, and Alex Zhavoronkov, "AI-powered therapeutic target discovery," *Trends in Pharmacological Sciences* 44, no 9 (September 2023): 561-572, https://pubmed.ncbi.nlm.nih.gov/37479540/; and Ece Kavalci and Anthony Hartshorn, "Improving Clinical Trial Design Using Interpretable Machine Learning Based Prediction of Early Trial Termination," *Scientific Reports* 13, no. 1 (January 4, 2023): 121, https://www.nature.com/articles/s41598-023-27416-7.

[20] Caroline Schuerger, Vikram Venkatram, and Katherine Quinn, *China and Medical AI: Implications of Big Biodata for the Bioeconomy* (CSET, May 2024), https://cset.georgetown.edu/publication/china-and-medical-ai/.

[21] *Charting the Future of Biotechnology* (Washington, D.C.: National Security Commission on Emerging Biotechnology, 2025), https://www.biotech.senate.gov/final-report/chapters/.

The 2025 Strategic Plan should prioritize research at the AI-biotech nexus, building on Strategies 1 and 8 from the 2023 Strategic Plan. Federal investment should focus on:

- Computational biology platforms integrating AI with experimental workflows;
- AI-driven drug discovery targeting areas underserved by commercial development, such as rare diseases and orphan conditions;[22]
- Secure, high-performance computing infrastructure specifically designed for sensitive biological data analysis;
- Interdisciplinary training programs combining AI expertise with life sciences.

**CNAS Intellectual Independence Statement**

As a research and policy institution committed to the highest standards of organizational, intellectual, and personal integrity, CNAS maintains strict intellectual independence and sole editorial direction and control over its ideas, projects, publications, events, and other research activities. CNAS does not take institutional positions on policy issues and the content of CNAS publications reflects the views of their authors alone. In keeping with its mission and values, CNAS does not engage in lobbying activity and complies fully with all applicable federal, state, and local laws. CNAS will not engage in any representational activities or advocacy on behalf of any entities or interests and, to the extent that the Center accepts funding from non-U.S. sources, its activities will be limited to bona fide scholastic, academic, and research-related activities, consistent with applicable federal law. The Center publicly acknowledges on its website annually all donors who contribute.

[22] Amit Gangwal and Antonio Lavecchia, "AI-Driven Drug Discovery for Rare Diseases," *Journal of Chemical Information and Modeling* 65, no. 5 (March 10, 2025): 2214-2231, https://doi.org/10.1021/acs.jcim.4c01966.