PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mba-6zto-786b
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-0332
Comment on FR Doc # 2025-07332

Submitter Information

Organization: Institute for AI Policy and Strategy

General Comment

IAPS respectfully submits its response to the OSTP RFI on a 2025 National AI R&D Strategic Plan. Please see attached file.

Docket ID No. NSF-2025-OGC-0001

Attachments

IAPS Submission to OSTP RFI on AI RD Strategic Plan

IAPS | Institute for AI Policy and Strategy

May 29, 2025

To: Suzanne Plimpton, RCO
National Science Foundation
2415 Eisenhower Avenue, Alexandria, VA 22314

Submitted electronically via Regulations.gov to Docket NSF-2025-OGC-0001

Response to OSTP RFI on the Development of a 2025 National Artificial Intelligence (AI) Research and Development (R&D) Strategic Plan

The Institute for AI Policy and Strategy respectfully submits its comments on the Office of Science and Technology Policy's Request for Information on the Development of a 2025 National Artificial Intelligence (AI) Research and Development (R&D) Strategic Plan. These comments focus on ways the AI R&D Strategic Plan can advance responsible AI innovation while maintaining America's technological leadership.

About the Institute for AI Policy and Strategy

The Institute for AI Policy and Strategy (IAPS) is a U.S.-based, nonpartisan policy research nonprofit. We engage experts across the U.S. and allied nations to deliver concrete, technically sound policy research that enhances national competitiveness and mitigates emerging risks while protecting the space for innovation to thrive. IAPS maintains strict intellectual independence and does not accept funding that could compromise the integrity of its research.

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.

Points of Contact: Joe O'Brien and Jam Kraprayoon , IAPS Researchers

Executive Summary

Artificial intelligence represents the most significant technological opportunity of our time, with potential to drive trillions in economic growth[1] and revolutionize scientific discovery.[2] However, this opportunity comes with substantial challenges. Foreign adversaries may leverage AI's capabilities to threaten U.S. national security and economic competitiveness. These adversaries will also likely target our AI research, supply chains, and infrastructure[3]. Beyond adversarial threats, as AI systems become more autonomous, they may become deceptive or even misaligned, creating additional risks.[4] The 2025 National AI R&D Strategic Plan offers a chance to guide federal research investments crucial for navigating these challenges and securing AI's benefits.

IAPS's work makes it uniquely positioned to recommend goals for the 2025 plan:

- **Research and Reports:** IAPS has published detailed reports covering infrastructure and compute, AI agents, and cybersecurity.
- **R&D Survey**: IAPS recently surveyed experts across 105 AI reliability and security research areas to identify promising research prospects for strategic AI R&D investment.[5]

Here we outline four research goals that the 2025 National AI R&D Strategic Plan should pursue:

I. Strengthen AI Model Security, Reliability, and Capability Assessments
II. Advance the Security of AI Supply Chains and Infrastructure
III. Develop Methods to Monitor Adversary AI Capabilities
IV. Evaluate, Secure, and Support Multi-Agent Systems

[1] The Epoch AI Team, "GATE: Modeling the Trajectory of AI and Automation," Epoch AI, March 21, 2025, https://epoch.ai/blog/announcing-gate.
[2] Dario Amodei, "Machines of Loving Grace," October 2024, https://www.darioamodei.com/essay/machines-of-loving-grace.
[3] Jeremie Harris and Edouard Harris, "America's Superintelligence Project," America's Superintelligence Project, April 2025, https://superintelligence.gladstone.ai/.
[4] Anthropic, "System Card: Claude Opus 4 & Claude Sonnet 4," May 2025, https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf.
[5] Joe O'Brien & Jeremy Dolan et al., "Expert Survey: AI Reliability & Security Research Priorities," Institute for AI Policy and Strategy, May 2025, https://www.iaps.ai/research/ai-reliability-survey.

Research Goal I: Strengthen AI Model Security, Reliability, and Capability Assessments

As AI capabilities scale and adoption accelerates, we must fund research ensuring model security and reliability to build systems that governments, businesses, and consumers can trust. **IAPS's expert survey identified high-importance but challenging areas requiring substantive research investments, primarily in AI security, including access control and interface hardening, supply chain integrity, weight security, and confidential computing.** The survey also flagged high-importance and high-tractability areas including capability evaluations for CBRN, cyber, and deception scenarios, as well as understanding emergence and scaling laws that govern how new capabilities develop.

Priority Research Areas

- **Understand emergence and task-specific scaling patterns**: This area involves formalizing and forecasting the emergence of new capabilities as models scale, investigating whether scaling alone can produce certain capabilities, and designing methods for discovering task-specific scaling laws.[6] This would improve preparedness and reduce the risk of strategic surprise.

- **Improve cybersecurity for AI models**: Cybersecurity research must protect model parameters, interfaces, and outputs from unauthorized access through

[6] Example work includes:
- Ian R. McKenzie et al., "Inverse Scaling: When Bigger Isn't Better" (arXiv, May 13, 2024), https://doi.org/10.48550/arXiv.2306.09479.https://doi.org/10.48550/arXiv.2306.09479
- Deep Ganguli et al., "Predictability and Surprise in Large Generative Models," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22 (New York, NY, USA: Association for Computing Machinery, 2022), 1747–64, https://doi.org/10.1145/3531146.3533229.https://doi.org/10.1145/3531146.3533229;
- Ethan Caballero et al., "Broken Neural Scaling Laws" (arXiv, July 24, 2023), https://doi.org/10.48550/arXiv.2210.14891.

cryptographic and architectural safeguards. Preventing data poisoning[7] or trojaning[8] from adversaries is also an issue. Work here includes ensuring secure weight storage, hardened access control, oracle protection measures, protecting algorithmic insights, preventing self-exfiltration, and robust data integrity.[9] This would prevent China or other adversaries from being able to steal our models, misuse our models, or tamper with them.

- **Advance domain-specific evaluations and evaluation science more broadly:** Domain-specific evaluation design requires specialized tools to assess AI capabilities in critical areas like automated AI research, cybersecurity, CBRN scenarios, and manipulative behaviors.Meanwhile, more fundamental research on the science of AI evaluations must be done to ensure that AI systems can be accurately assessed and understood.[10] This would help improve AI reliability and increase AI adoption.

- **Drive research highly neglected by industry:** One of our reports conducted a literature review to identify gaps in research left by industry.[11] We found that multi-agent safety, model organisms of misalignment, unlearning, safety-by-design,

[7] Bart Lenaerts-Bergmans, "What Is Data Poisoning? | CrowdStrike," CrowdStrike.com, March 19, 2024, https://www.crowdstrike.com/en-us/cybersecurity-101/cyberattacks/data-poisoning/.
[8] NIST, "What Is TrojAI," accessed May 29, 2025, https://pages.nist.gov/trojai/docs/about.html.
[9] Example work includes:
- Sella Nevo et al., "Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models," May 30, 2024, https://www.rand.org/pubs/research_reports/RRA2849-1.html.
- Joshua Clymer, Hjalmar Wijk, and Beth Barnes, "The Rogue Replication Threat Model," *METR Blog*, November 12, 2024, https://metr.org/blog/2024-11-12-rogue-replication-threat-model/.

[10] Example work includes:
- Hjalmar Wijk et al., "RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents against Human Experts" (arXiv, May 27, 2025), https://doi.org/10.48550/arXiv.2411.15114.
- Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn, "Large Language Models Can Strategically Deceive Their Users When Put Under Pressure" (arXiv, July 15, 2024), https://doi.org/10.48550/arXiv.2311.07590.

[11] Oscar Delaney, Oliver Guest, and Zoe Williams, "Mapping Technical Safety Research at AI Companies: A Literature Review and Incentives Analysis" (arXiv, September 25, 2024), https://doi.org/10.48550/arXiv.2409.07878.

and controlling untrusted AIs[12] were particularly neglected in terms of public industry research. Closing these gaps would improve AI reliability and security.

Research Goal II: Advance the Security of AI Supply Chains and Infrastructure

Deep AI integration across sensitive applications requires verifiably secure infrastructure that addresses critical vulnerabilities in data centers, including insider threats and supply chain attacks. Securing AI systems at the hardware level involves protecting model weights, securing deployment environments, maintaining supply chain integrity, and implementing robust monitoring through confidential computing, rigorous access controls, specialized hardware protections, and continuous security oversight.[13] Even high-standard facilities face significant risks from nation-states, and end users lack reliable methods to continuously verify security status.

IAPS's expert survey identified supply chain integrity, weight security, and confidential computing as areas requiring more substantive investments of time and research. U.S. government can uniquely address the need to develop novel technologies including efficient verification of hardware designs, tamper-resistant chip hardening, and scalable tamper-evidence techniques. As AI is a key strategic technology, these techniques need to be robust against heavily resourced nation-state actors – a level of security that industry is likely not to be able to meet without help from the U.S. government. Priority research areas include remotely verifiable tamper-evidence like physical unclonable functions, remote attestation protocols for collective integrity verification, and enhanced authentication techniques to detect supply chain attacks through large-scale chip scanning for implants or modifications.

[12] We suspect that controlling untrusted AIs is now less neglected by AI companies than at the time of writing the report.

[13] Example work includes:

- Nevo et al., "Securing AI Model Weights."
- Isaac Hepworth et al., "Securing the AI Software Supply Chain," 2024, https://research.google/pubs/securing-the-ai-software-supply-chain/.

Priority Research Areas

- **Hardware security to protect against insider threats and supply chain attacks**, including:
    - Methods to more efficiently verify that chip designs are free from vulnerabilities, including vulnerabilities to side channel attacks.[14]
    - Methods to harden chips against tampering, including high-security, scalable tamper-resistance and tamper-evidence techniques.[15]

- **Improved hardware and software capabilities and protocols for verifying the state of heterogeneous data center systems**, including:
    - Remotely verifiable high-security tamper-evidence, such as enclosures that form physical unclonable functions (PUFs).[16]
    - Remote attestation and encryption protocols to enable heterogeneous devices to collectively attest to the integrity and secure configuration and form verifiably encrypted enclaves.[17]

- **Improved techniques for authenticating components to detect supply chain attacks**, including scanning chips at scale to detect implants or modifications.[18]

---

[14] Example work includes: Watson, Robert N M, Peter G Neumann, Jonathan Woodruff, Michael Roe, Hesham Almatary, Jonathan Anderson, John Baldwin, et al. "Capability Hardware Enhanced RISC Instructions: CHERI Instruction-Set Architecture (Version 7)." University of Cambridge Computer Laboratory, June 2019.

[15] Example work includes: Mosavirik, Tahoura, and Shahin Tajik. "IC Backside Tamper Detection Using Impedance Sensing." *IEEE Access* 13 (2025): 90706–25. https://doi.org/10.1109/ACCESS.2025.3572694.

[16] Example work includes: Immler, Vincent, Johannes Obermaier, Kuan Kuan Ng, Fei Xiang Ke, JinYu Lee, Yak Peng Lim, Wei Koon Oh, Keng Hoong Wee, and Georg Sigl. "Secure Physical Enclosures from Covers with Tamper-Resistance." *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2019, 51–96. https://doi.org/10.13154/tches.v2019.i1.51-96.

[17] Example work includes: Zhu, Jianping, Rui Hou, XiaoFeng Wang, Wenhao Wang, Jiangfeng Cao, Boyan Zhao, Zhongpu Wang, et al. "Enabling Rack-Scale Confidential Computing Using Heterogeneous Trusted Execution Environment." In *2020 IEEE Symposium on Security and Privacy (SP)*, 1450–65. San Francisco, CA, USA: IEEE, 2020. https://doi.org/10.1109/SP40000.2020.00054.

[18] Example work includes: Moore, Samuel K. "X-Ray Tech Lays Chip Secrets Bare." IEEE Spectrum, October 7, 2019. https://spectrum.ieee.org/xray-tech-lays-chip-secrets-bare.

Research Goal III: Develop Methods to Monitor Adversary AI Capabilities

While operational intelligence may be outside of the strategic plan's scope, the plan can fund basic science that *underpins* monitoring technologies. U.S. economic and national security requires detailed intelligence on adversary AI development efforts to prevent strategic surprise from competitors developing advanced AI systems with significant security implications.

While the U.S. government likely possesses sophisticated classified monitoring capabilities, OSTP should conduct a comprehensive review to identify and address critical gaps in existing systems. The Federal Government must invest in R&D to enhance its ability to track adversary AI progress, particularly as nations like China attempt to circumvent U.S.-led export controls through innovative approaches to AI development.

Priority research areas include understanding how adversaries might leverage geographically distributed compute clusters for model training, detecting large cloud workloads used for AI development, and developing remote monitoring methods to identify clandestine datacenters through satellite imagery, energy consumption analysis, and financial transaction monitoring. These monitoring techniques require rigorous real-world testing and refinement, including validation against known datacenters and red team evasion attempts. Once operational, the U.S. must develop robust counterintelligence measures to protect American AI development from foreign surveillance, addressing high-importance but challenging areas like access control, supply chain integrity, and confidential computing that currently face significant tractability constraints.

Priority Research Areas

- **Understanding how adversaries could leverage geographically disparate compute clusters for model training:** In an attempt to train powerful AIs with its limited supply of AI chips - partly a result of U.S.-led export controls - the People's Republic of China (PRC) might attempt to combine multiple existing AI clusters. U.S. government policy responses to such an effort would benefit from a detailed

understanding of how capable the resulting PRC AIs are likely to be. Reuel, Bucknall, et al. (2024) call this an open problem; see Section 3.2.1.[19]

- **Detecting large cloud compute workloads used for model training:** Leveraging cloud compute is another avenue that geopolitical rivals may use to train large AI models, but technology does not currently exist to reliably detect how those resources are being used. Reuel, Bucknall, et al. (2024) call this an open problem; see Sections 3.2.2 and 5.2.2.[20]

- **Identifying remote monitoring methods to detect clandestine datacenters or mobile compute clusters:** Such methods would make it harder for adversaries to train powerful AIs in secret. Examples of remote monitoring methods could include remote sensing (such as satellite imagery to spot construction footprints), monitoring energy consumption from power grids, and monitoring suspicious financial transactions. Machine learning tools might facilitate the processing of large amounts of data from these methods. Wasil et al. (2024) give an overview of these methods but further work is needed to identify specific avenues for improving them with R&D.[21]

Research Goal IV: Evaluate, Secure, and Support Multi-Agent Systems

As generative AI enables autonomous agents to execute complex multi-step tasks[22], the shift from isolated systems to multi-agent interactions introduces critical new challenges that demand urgent research attention. These systems exhibit emergent behaviors, systematic instabilities that can cascade into failures, and expanded attack surfaces that

[19] Anka Reuel et al., "Open Problems in Technical AI Governance" (arXiv, April 16, 2025), https://doi.org/10.48550/arXiv.2407.14981.
[20] Reuel et al.
[21] Akash R. Wasil et al., "Verification Methods for International AI Agreements" (arXiv, November 4, 2024), https://doi.org/10.48550/arXiv.2408.16074.
[22] Jam Kraprayoon et al., "AI Agent Governance: A Field Guide," Institute for AI Policy and Strategy, April 17, 2025, https://www.iaps.ai/research/ai-agent-governance.

malicious actors can exploit.[23] **IAPS's expert survey confirms multi-agent systems as a top priority, with all multi-agent related research areas ranking in the top 30. Experts identified immediate focus areas, including metrics, oversight, and monitoring.**[24]

The U.S. government has a strategic opportunity to drive fundamental research that prepares society for widespread AI agent deployment, as inadequate security investments could stall industry adoption and cause America to forfeit significant productivity gains—similar to how security concerns have slowed Internet of Things adoption.[25] This research must develop a deeper understanding of how LLM-based agents learn, respond to underspecified goals, and interact with their environments. Federal investment should also support research that secures multi-agent interactions through detection of harmful collective behaviors, transparency studies, and robust evaluation frameworks for agent dynamics.[26]

Priority Research Areas

- **Multi-agent evaluations and metrics:** Evaluations of AI agents in isolation are insufficient to understand or predict the complex, potentially harmful, emergent behaviors and vulnerabilities that arise in multi-agent settings. While there is some nascent work on multi-agent-relevant evaluations[27], there are still no well-validated definitions, metrics, and methods for many multi-agent capabilities and propensities.

[23] Lewis Hammond et al., "Multi-Agent Risks from Advanced AI," February 19, 2025, https://arxiv.org/abs/2502.14143.
[24] O'Brien & Dolan et al., "Expert Survey."
[25] Katerina Megas, "IoT Assignment Completed! Report on Barriers to U.S. IoT Adoption," *NIST*, October 22, 2024, https://www.nist.gov/blogs/cybersecurity-insights/iot-assignment-completed-report-barriers-us-iot-adoption.
[26] Example work includes:
- Silen Naihin et al., "Testing Language Model Agents Safely in the Wild" (arXiv, December 3, 2023), https://doi.org/10.48550/arXiv.2311.10538.
- Donghyun Lee and Mo Tiwari, "Prompt Infection: LLM-to-LLM Prompt Injection within Multi-Agent Systems" (arXiv, October 9, 2024), https://doi.org/10.48550/arXiv.2410.07283.

[27] E.g., Sumeet Ramesh Motwani et al., "Secret Collusion among Generative AI Agents: Multi-Agent Deception via Steganography" (arXiv, April 14, 2025), https://doi.org/10.48550/arXiv.2402.07510.

- **Secure interaction protocols and environments:** Currently, the communication pathways between frontier AI agents are underspecified, even though they are a central source of risk from multi-agent systems. Unsecured, free-form communications between agents allow for agent-to-agent attacks or secret collusion. Interaction standards need to be developed that embed security, privacy, and governance, drawing potentially from developments in secure multi-party computation and verifiable interactions.[28]

- **Monitoring and oversight systems:** AI agents, especially when deployed en masse, will have a volume and speed of actions that will be exceedingly difficult for humans to track and check manually. Given this, there should be more research effort focused on developing agent-compatible monitoring and oversight systems, such as those that employ automated oversight[29] and robust threat attribution mechanisms.

Conclusion

By prioritizing these four goals, the federal government can address critical technical challenges that industry cannot solve alone, while creating the foundation for safe, secure, reliable, and beneficial AI deployment at scale. Proactive research into high-priority AI reliability and security areas, infrastructure security, global AI landscape awareness, and safe multi-agent coordination will enable innovation while preventing the need for reactive measures that could stifle progress.

[28] Christian Schroeder de Witt, "Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents" (arXiv, May 4, 2025), https://doi.org/10.48550/arXiv.2505.02077.
[29] Naihin et al., "Testing Language Model Agents Safely in the Wild."