IAPS | Institute for AI Policy and Strategy

March 14, 2025

To: Faisal D'Souza, NCO
Office of Science and Technology Policy
2415 Eisenhower Avenue, Alexandria, VA 22314

Submitted by email to [redacted]

Response to OSTP RFI on AI Action Plan

Docket ID: 90 FR 9088, NSF_FRDOC_0001

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 an Artificial Intelligence (AI) Action Plan. Our comments focus on ways the AI Action Plan can build trust in American AI, deny advantages to adversaries, and prepare to adapt as the technology evolves.

About the Institute for AI Policy and Strategy

The Institute for AI Policy and Strategy (IAPS) is a 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 AI Action Plan and associated documents without attribution.

Point of Contact: Jenny Marron, Director of Policy and Engagement ([redacted]

Executive Summary

The U.S. is the global leader in AI development, home to most of the world's leading advanced foundation models. U.S. companies like NVIDIA, Microsoft, and Amazon are leading players in building cutting-edge AI infrastructure, and American start-ups like Scale AI and Databricks have rapidly innovated across the AI value chain.

America's leadership, however, is under threat. Due to insufficient earlier export controls, China is advancing rapidly. Chinese entities like DeepSeek have developed frontier models that show they are only about six months behind leading American AI systems. America must not cede territory with AI like it has for other critical technologies like hypersonics and 5G.

AI is a key strategic technology for the U.S., most notably because AI could eventually exceed human capabilities in a wide variety of relevant domains for defense. While experts differ on the timeline, we are likely to see continued major breakthroughs during the current Administration. We share the President's vision for seeing AI as an opportunity for America. Advancements in AI bring enormous potential for scientific, economic, and productivity gains and the benefits for Americans could be tremendous. However, we must also be mindful of the risks – future, more advanced AI may produce domestic market disruptions and the rise of weaponized AI attacks from U.S. adversaries. These disruptions would have profound implications for national security, geopolitical stability, and the everyday life of American citizens. The public will expect the government to understand and manage these disruptions.

To meet the Trump Administration's stated policy to "sustain and enhance America's global AI dominance in order to promote human flourishing, economic competitiveness, and national security," the AI Action Plan should outline steps that secure economic growth and prosperity for its citizens and retain a strategic advantage against foreign adversaries. The U.S. federal government should provide strategic technical leadership on AI through focused expertise that maximizes America's competitive edge. By emphasizing specialized capabilities in targeted areas – rather than headcount – the government can excel in the areas where it is uniquely positioned to lead relative to the market: addressing national security challenges, supporting fundamental research, and establishing standards.

We recommended the AI Action Plan include three key areas:

- **Build trust in American AI**: Establish AI systems that governments, businesses, and consumers can trust through enhanced security and reliability standards. Leverage federal capabilities to address critical market gaps and secure AI supply chains against malicious disruption.

- **Deny foreign adversary access to advanced computing technology:** Maintain America's technological advantage by controlling semiconductor exports to adversaries, forcing them to choose between research advancement and deployment. Coordinate across government agencies to ensure effective implementation of these controls.

- **Understand and respond to changing capabilities**: The United States needs the ability and agility to respond at speed as technology evolves. By developing the systems and standards now, the Administration creates optionality for responding in the future. Develop evaluation standards to assess emerging AI systems and their national security implications. Create coordinated visibility across government, industry, and research institutions to promote beneficial AI while addressing security concerns.

Detailed recommendations

Goal I: Build trust in American AI

For American AI to transform the world, it must first earn the trust of governments, businesses, and consumers. Systems that diagnose diseases, offer autonomous transportation, and manage critical infrastructure must be both secure and reliable. From aviation protocols to encryption standards, the U.S. government has repeatedly pioneered research and frameworks later adopted throughout industry that has enabled innovation to thrive. By strategically addressing gaps in private sector research and investment, particularly in areas like AI security assurance and reliability testing, federal initiatives can provide significant encouragement for consumer adoption. The federal government should also leverage its unique capabilities and authorities to secure AI and advanced computing supply chains to prevent illicit adversary theft and/or tampering, undermining American competitiveness and security. As American innovation accelerates and export controls restrict adversarial access, foreign actors will increasingly target private sector AI assets and infrastructure. Model theft, data poisoning, and model trojans remain key threats.

1.1 Leverage R&D and Standards Development to Ensure American AI Systems are Secure and Reliable

Targeted government efforts can fill important gaps in AI research that private companies overlook or open-source developers need, particularly in areas like evaluation science, multi-agent interaction, and model security. American leadership in developing and promoting technical standards is also essential for national security and economic competitiveness. Foreign adversaries are actively working to influence emerging technology standards through strategic initiatives like Standards 2035, having already attempted to undermine U.S. standards in telecommunications and quantum encryption. Other federal efforts, such as tracking software vulnerabilities, help developers quickly identify and correct issues. These programs need modernization to address the unique challenges posed by AI-related vulnerabilities.

Advance AI Security and Assurance Technology

- Direct federal civilian and defense research agencies to prioritize funding research that helps improve the security and reliability of AI models. Agencies should leverage unique authorities to accelerate research, promote competitive research, and collaborate with nontraditional contractors.
    - OSTP, with support from OMB, should include a list of critical AI security technologies in vehicles such as the annual multi-agency R&D priorities memoranda and the next update of the National R&D Strategic Plan, as well as work with AI R&D funders to develop technology roadmaps that detail related technical benchmarks and milestones, capability development timelines, resource requirements, and stakeholder roles and responsibilities.
    - Priority research areas are summarized in the table below.

Recommended priority areas for AI security and assurance R&D[1]

R&D areas	Description
Hardware and infrastructure security	Ensuring the security of AI systems at the hardware and infrastructure level involves protecting model weights, securing deployment environments, maintaining supply chain integrity, and implementing robust monitoring and threat detection mechanisms. Methods include the use of confidential computing, rigorous access controls, specialized hardware protections, and continuous security oversight. Example work includes Nevo et al. (2024) and Hepworth et al. (2024)
Agent safety and multi-agent interaction	Developing a deeper understanding of agentic behavior in LLM-based systems, including clarifying how LLM agents learn over time, respond to underspecified goals, and engage with their environments. This also includes research focusing on ensuring safe multi-agent interactions, such as by detecting and preventing malicious collective behaviors, studying how transparency can affect agent interactions, and developing evaluations for agent behavior and interaction. Example work includes Naihin et al. (2023) and Lee & Tiwari (2024)
Cybersecurity for AI models	Focusing on protecting model parameters, interfaces, training techniques, and outputs from unauthorized access, extraction, or misuse using cryptographic, architectural, and procedural safeguards. This includes ensuring secure weight storage, hardened access control, oracle protection measures, protecting algorithmic insights, preventing self-exfiltration, and robust data integrity. Example work includes Nevo et al. (2024) and Clymer et al. (2024)
Domain-specific AI evaluation design and improving evaluation science	Developing specialized evaluation tools to assess AI models' capabilities and safety in critical areas such as automated AI research and development, cybersecurity, chemical/biological/radiological/nuclear (CBRN) scenarios, and manipulative behaviors like deception and persuasion. This also includes broader research on AI evaluations to ensure that, generally, AI systems can be accurately assessed and understood. This includes theoretical work in capability and safety evaluation and improving the reliability and fairness of evaluation processes. Example work includes Wijk et al. (2024) and Scheurer et al. (2023)
Understanding in-context learning, reasoning, and scaling behavior	Methods to gain a comprehensive understanding of how large language models learn, reason, and scale, such as by examining in-context learning (ICL) mechanisms, the influence of data and design on behavior, the theoretical foundations of scaling, the emergence of advanced capabilities, and the nature of reasoning. Example work includes Olsson et al. (2022) and McKenzie et al. (2023)

[1] IAPS has conducted research to identify priority AI assurance and security R&D areas, see Delaney et al. 2024; Kraprayoon and Anderson-Samways 2024; and O'Brien et al. *forthcoming*.

Establish Federal AI Research Initiatives and Infrastructure

- Establish dedicated research centers within DOE National Laboratories focused on improving AI system security and reliability. Areas of research should include explainability, secure architectures, and adversarial resilience.
- Invest in secure computing infrastructure and classified test environments to rigorously assess AI systems under simulated adversarial conditions.
- Provide U.S. researchers and academics with access to public computational, data, and training resources. This should include providing ongoing support and funding to the National AI Research Resource (NAIRR).

Develop AI Assurance Standards and Guidance for Development and Deployment

- Direct NIST, in coordination with CISA and NSA, to develop comprehensive standards for securing AI systems, including guidance on secure development practices (i.e. NIST SP 800-218A), vulnerability management in models and scaffolding, deployment configurations, and AI agent-specific security controls.
- Direct NIST to develop standards and guidance for AI system reliability, focusing on reliable design methodologies, robust testing frameworks, and operational deployment considerations to ensure consistent performance and accuracy across varied production environments.
- Direct sector-specific agencies, in coordination with NIST, to develop tailored AI reliability guidelines addressing unique operational requirements, risk profiles, and compliance considerations for their respective industries.

Strengthen AI Security Vulnerability Tracking and Disclosure

- Direct CISA to either update the Common Vulnerabilities and Exposure (CVE) program or develop a new process specifically designed to track and catalog AI security vulnerabilities, improving the identification and mitigation of AI-related cybersecurity threats.
- Direct NIST to update the National Vulnerability Database (NVD) to better accommodate and categorize AI-specific vulnerabilities, enhancing the repository's ability to serve as a comprehensive resource for AI security risks.

1.2 Secure America's AI and Advanced Computing Supply Chain

The AI and advanced computing supply chain is crucial for both building trust in AI systems and maintaining America's lead over adversaries. As AI capabilities advance and U.S. semiconductor export controls slow adversarial innovation, America's AI and advanced computing industries will become an increasingly attractive target. For

example, by gaining access to unreleased models, hostile nation-states could acquire advanced capabilities at a fraction of the cost, sabotage AI systems, or accelerate their own R&D. Further, nation-state cyber and espionage operations are growing more common, capable, and strategic, potentially surpassing what even the most well-resourced companies can effectively counter alone. Only governments possess the unique authorities, intelligence capabilities, and cross-sector coordination essential for protecting these strategic national assets from both compromise and disruption.

Define and Advance Security Standards for AI Model Weights and Other Critical Assets

- Direct NIST to develop security standards for model weights (equivalent to SL4 and SL5 as outlined by the RAND Corporation) and other critical assets beyond model weights (i.e. algorithms and training data).
- Prioritize research that supports the development of technologies required to meet or exceed the SL4 and SL5 security standards.

Secure the AI and Advanced Computing Supply Chain from Adversarial Tampering and Distribution

- Direct relevant agencies to expand AI security research efforts and establish competitive initiatives to prevent model sabotage and tampering, for example by broadening IARPA's TrojAI program to include comprehensive defensive controls and launching cross-sector Red Team R&D programs that perform adversarial testing throughout the AI model lifecycle.
- Direct relevant agencies to strengthen AI supply chain security and resilience by taking actions such as identifying critical hardware components for domestic production, evaluating the AI software supply chain for vulnerabilities, assessing risks to critical nodes, and sharing supply chain risk information.

Secure the AI and Advanced Computing Sector

- Designate AI and Advanced Computing (AIAC) as a critical infrastructure sector. The sector should include stakeholders in the AI supply chain (i.e. AI developers, cloud hyperscalers, semiconductors manufacturers).
- Designate DHS as the SRMA for the AIAC sector to provide services, technical assistance, and coordinated public-private collaboration efforts.
- Direct the intelligence community to prioritize identifying and analyzing nation-state efforts to target the AIAC sector.

Improve Threat Information Sharing

- Pilot a public-private cybersecurity information sharing program, similar to DOE's CRISP or CISA's CyberSentry, for AI developers. If successful, this program could scale to other AI and Advancing Computing providers.
- Provide AI developers and Advanced Computing providers with access to classified cyber threat information and briefings. This collaboration could be modeled after programs like the ODNI's Critical Infrastructure Intelligence Initiative.

1.3 | Strengthen Government's Ability to Drive AI Innovation and Assurance

To be an effective partner to industry, Federal Government agencies need clear roles, specialized expertise, and dedicated resources. The private sector should not have to navigate a byzantine and uncoordinated maze of government agencies to find support. To drive economic benefits, sector-specific agencies also need the expertise to understand the unique opportunities and challenges within their domains, helping their sectors safely deploy AI by providing tailored guidance and removing regulatory barriers.

Determine Federal Roles and Responsibilities

- Issue a White House policy directive that identifies and clarifies Federal agencies' roles and responsibilities related to AI and advanced computing. The directive should establish lead and supporting roles to address AI policy issues, including AI evaluations, standards development, and supply chain security. This should include designating a primary federal government point of contact with private sector AI developers to facilitate voluntary testing of dual-use foundation models.

Establish a US AI Center of Excellence (USAICoE)

- Establish a centralized node to enable AI use by evaluating emerging AI capabilities, developing assurance standards, and fostering close collaboration with industry. For the purposes of this RFI, we will refer to it as a federal AI Center of Excellence within NIST. Key functions should include:
    - Advancing AI measurement and evaluation science, providing both the private and public sector with the tools to identify and understand AI's economic opportunities and potentially dangerous capabilities.
    - Conducting technical evaluations by working with AI developers.
    - Developing and promoting standards and guidance, including assurance standards that improve the security and reliability of AI systems.

    - Serving as a source of expertise and coordinating with other Federal Agencies, including helping sector-specific agencies promote safe AI deployment within their respective sectors.
    - Engaging with external stakeholders, including AI developers, Standards Development Organizations (SDOs), and the AI institutes of other countries.
- This could be accomplished by restructuring, re-housing, or replacing the US AI Safety Institute (US AISI), but it is critical that the U.S. government have a center of gravity for these functions.

Establish Sector-Specific AI Innovation and Assurance Hubs

- Establish Sector-Specific AI Innovation and Assurance Hubs within existing departments or agencies. These hubs could promote safe AI deployment within their respective sectors. Functions may include helping industry find innovative uses, supporting adoption with pilot programs that remove regulatory barriers, and developing tailored assurance guidance for sector-specific applications. Through a whole-of-government approach, NIST's Federal AI Center of Excellence would support the sector-specific hubs, providing general AI technical expertise.

Goal II: Deny foreign adversary access to advanced computing technology

Maintaining America's technological advantage in AI and advanced computing is essential for national security. Scaling laws and increasing computational demands for inference mean semiconductors and related technology will remain crucial for AI advancement. The impacts of effective semiconductor export controls will compound over time, slowing adversarial research and deployment. However, export control enforcement requires a whole-of-government approach. The Bureau of Industry and Security (BIS) cannot single-handedly counter smuggling, identify technical loopholes, and lead international coalitions. Effective implementation demands coordinated action across the intelligence community, the State Department, the Department of Homeland Security, and technical agencies like NIST.

2.1 | Prevent Foreign Adversaries' Access to U.S. Advanced Computing Technology

Effective implementation of export controls requires addressing smuggling, closing loopholes, deploying innovative technology and, most critically, staying the course. For

example, DeepSeek's recent success resulted from insufficient controls established in 2022, not from failures in the current approach. If the administration can strengthen controls, the compounding effects will significantly constrain adversarial AI development. It's already estimated that AI companies allocate 60-80% of their compute resources to deployment. This, combined with reasoning models inference time compute demands, means adversaries will be forced to choose between research and deployment. And unlike the delayed impacts on development, export controls will have an immediate impact on deployment.

Strengthen Export Controls and Enforcement

- Establish a Joint Federal Task Force, led by a revitalized BIS, focused on stopping the diversion of AI chips and illegal tech transfer of information relevant to advanced AI semiconductor manufacturing, such as electronic design automation (EDA) software piracy. The administration should use all relevant policy tools and authorities to enforce semiconductor export controls. The Task Force should include ODNI, DOJ, State, and DHS, and prioritize improved interagency coordination between the IC and BIS.
- Direct ODNI to collect and share relevant intelligence with BIS to strengthen export control enforcement, including through mapping smuggling networks and weak points in the AI chip distribution network. This would enable BIS to target enforcement more efficiently.
- Direct NIST to collaborate with industry to identify hardware security features and other technology that can support export control enforcement and deter smuggling. This should include commissioning a feasibility study of delay-based location verification for AI chips and creating a centralized chip registry pilot database within BIS. These features could enable more efficient enforcement generally and allow industry to export to higher-risk destinations, such as the Middle East, while reducing the risk of chips being smuggled to China.[2]
- Direct BIS to expand export controls to include NVIDIA H20 chips and equivalents, while also reviewing whether some consumer GPUs need to be more strongly controlled.[3]
- Establish a BIS whistleblower program to incentivize reports of export violations, funded via penalties levied on violators.

[2] For more information on these technologies, see Asher Brass and Onni Aarne, "Location Verification for AI Chips," Institute for AI Policy and Strategy April 2024, https://www.iaps.ai/research/location-verification-for-ai-chips, and Onni Aarne, Tim Fist, and Caleb Withers, "Secure, Governable Chips," Center for New American Security, January 8, 2024, https://www.cnas.org/publications/reports/secure-governable-chips.

[3] See also Erich Grunewald, "Are consumer GPUs a problem for US Export Controls?", May 2024, https://www.iaps.ai/research/are-consumer-gpus-a-problem-for-us-export-controls.

Preserve America's Compute Advantage

- Direct relevant federal agencies to collaborate with industry, including AI infrastructure providers operating overseas data centers, to create a strategy for securing offshore AI infrastructure against foreign cyber operations. This strategy should identify baseline security requirements, federal support efforts, and recommendations for Congress and the President.
- Revise the AI diffusion rule to create clear criteria for countries to gain 'Tier 1' status, e.g. by improving their export control enforcement practices, to ensure the right set of countries are in the Tier 1 group and incentivize Tier 2 countries to better enact controls.
- Direct DOC to establish reporting requirements for cloud computing providers regarding sales metrics and transaction details with Chinese entities, including customer verification procedures and compliance with export control.

Goal III: Understand and Respond to Changing Capabilities

As AI capabilities advance, the federal government needs methods to assess emerging capabilities and understand their potential implications for national security. Developing rigorous evaluation frameworks and measurement standards enables identification of dual-use applications before they present unanticipated risks. Through coordinated efforts between government agencies, industry partners, and research institutions, America can maintain comprehensive awareness of both domestic innovations and foreign developments. This improved visibility gives decision-makers the insights needed to promote beneficial AI advancement while addressing genuine security concerns, avoiding unnecessary regulation and unseen risks. The United States should develop these capabilities and systems now - before they are needed - in order to have the optionality and agility to respond as the situation changes.

3.1 Advance AI Measurement and Evaluation Standards

Many experts believe AI capabilities will dramatically improve in the next few years. However, AI evaluation science is still in its infancy, lacking the scientific rigor needed to accurately assess rapidly emerging capabilities. Without better evaluation methods, both industry and government will navigate the road ahead in the dark. By helping create scientifically robust and government-backed evaluation standards, the US government can improve its own decision-making, help industry, and promote a third-party evaluation ecosystem.

Develop AI Evaluation Standards and Guidance

- Direct NIST to update AI capability evaluation standards and guidance to handle a broader range of national-security relevant capabilities beyond cybersecurity and CBRN. These standards should be based on the latest measurement science and updated frequently to keep pace with emerging AI system capabilities.
- Additional key capability areas for evaluations include:
    - Agent and multi-agent interactions (e.g., collusion capability between agents)
    - Model deception, scheming, and situational awareness
    - Automated AI research and development capabilities

Enable Private Sector and Third-Party Evaluators

- Direct NIST or other relevant federal agencies to provide guidance that helps companies encourage independent third-party testing of AI systems. This should cover both traditional security vulnerabilities and AI-specific risks that may lead to malicious use. This could include guidance on vulnerability disclosure programs and bug bounty initiatives that protect good-faith researchers from liability, such as rules of engagement that define testing boundaries, permitted methods, and reporting procedures.

3.2 Monitor and Assess AI for National Security Implications

Advanced AI systems pose significant national security risks if deployed by malicious actors or foreign adversaries. For example, cybersecurity researchers have already created AI systems that can identify zero-day vulnerabilities and conduct complex multi-stage attacks. Furthermore, OpenAI and Anthropic have both indicated in their latest model system cards that models that will be released later this year likely will be capable of guiding novices through launching known bioweapon and chemical weapon attacks[4]. Visibility into these emerging dual-use capabilities and foreign adversarial developments is imperative to both effectively mitigate risks and avoid unnecessary

[4] See OpenAI's Deep Research model system card (p17): "Several of our biology evaluations indicate our models are on the cusp of being able to meaningfully help novices create known biological threats" and Anthropic's Claude 3.7 Sonnet model system card (p24): "However, the results from our evaluations suggest improved performance in all domains, including some uplift in CBRN evaluations. [...] Further, based on what we observed in our recent CBRN testing, we believe there is a substantial probability that our next model may require ASL-3 safeguards", safeguards meant for working with models "increase the risk of catastrophic misuse compared to non-AI baselines (e.g. search engines or textbooks)".

regulation. This challenge demands close collaboration with industry and robust government evaluation capabilities.[5]

Identify National Security Relevant AI Capabilities

- Direct the USAICoE, in coordination with all relevant federal agencies, to lead evaluation efforts to identify emerging frontier model capabilities that could support or threaten US national security. This should include both classified (confidential) and unclassified (publicly available) evaluations. When appropriate, the evaluating agencies should enter into agreements with model developers to receive early access and provide feedback. The administration should consider tasking specific agencies with developing domain specific evaluations with the USAICoE supporting.This could include:
    - Direct the NSA to develop classified offensive and defensive cyber capabilities evaluations.
    - Direct the National Nuclear Security Administration (NNSA) to develop classified evaluations of nuclear and radiological relevant capabilities.
    - Direct USAICoE, in coordination with DHS and HHS, and other relevant agencies to develop classified evaluations of capabilities that could generate or exacerbate chemical and biological risks.

Monitor and Assess Foreign Adversary AI Capabilities

- Direct ODNI to assess strategic adversaries' AI capabilities, examining talent flows, computing resources, leadership intentions, and contingency measures for restricting adversarial AI systems when required. Classified findings should be submitted to the White House, China Select Committee, and Senate Intelligence Committee.

Develop Agile Systems to Identify and Respond to Emerging Risks

- The White House should establish a Rapid Emerging Assessment Council for Threats (REACT), able to rapidly convene cross-disciplinary subject matter experts to assess sudden, emerging, or novel AI-related threats to critical infrastructure or national security where government, industry, and academia may need to convene quickly to understand and mitigate sudden risks.
- NIST and the US Army Intelligence Center of Excellence should maintain the Testing Risks of AI for National Security (TRAINS) Taskforce and assign agency

[5] For more detail, see Joe O'Brien, Shaun Ee, Jam Kraprayoon, Bill Anderson-Samways, Oscar Delaney, and Zoe Williams "Coordinated Disclosure of Dual-Use Capabilities: An Early Warning System for Advanced AI," June 2024. https://www.iaps.ai/research/coordinated-disclosure

leads that will coordinate responses to reports of national security and public safety-relevant capabilities in frontier AI systems that arise from testing.
- NIST should solicit input on definitions, procedures, best practices, and guidelines for reporting and documentation of security and security-critical information about frontier AI systems.

Conclusion

The Administration has the opportunity through the AI Action Plan to set out a vision for AI that is secure, reliable, and able to achieve the promise of transformative economic and societal gains. The Institute for AI Policy and Strategy is grateful for the opportunity to offer recommendations for the federal government's capabilities to build trustworthy AI systems, deny adversaries access to advanced computing, and develop agile response mechanisms to emerging threats. This balanced approach—supporting industry leadership while fulfilling the government's fundamental obligation to defend its citizens—will ensure America maintains its competitive edge in the AI race and harnesses these powerful technologies to enhance national security and economic prosperity. We welcome the opportunity to answer any of your questions or engage in more detail as OSTP considers these recommendations.