PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mb9-wnp7-nhuh
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-0271
Comment on FR Doc # 2025-07332

Submitter Information

Organization: Machine Intelligence Research Institute

General Comment

Please see the attached file for our response to the RFI on the Development of a 2025 National AI R&D Strategic Plan.

Attachments

MIRI TGT RFI on National AI Research and Development Strategic Plan

Peter Barnett, David Abecassis, Aaron Scher
Technical Governance Team
Machine Intelligence Research Institute
Berkeley CA
techgov@intelligence.org

May 29, 2025

Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO), National Science Foundation

Re: Development of a 2025 National AI R&D Strategic Plan. Request for Information.

The Machine Intelligence Research Institute (MIRI) is a nonprofit based in Berkeley, California, founded in 2000. For over two decades, MIRI has worked to understand and prepare for the critical challenges that humanity will face as it transitions to a world with artificial superintelligence.

In our view, the field of AI is on a course to build AI systems that can substantially surpass humanity in all strategically relevant activities (often referred to as "artificial superintelligence"), with little to no understanding of how they function or ability to robustly steer and control their behavior. We and many other experts believe the development of such systems could cause human extinction. To avoid this risk, we believe it will be necessary to put in place coordinated global checks on AI development, capable of enabling, enforcing, and verifying restrictions on the development of superintelligent AI systems, until such a time as humanity's understanding of the relevant systems is sufficient to ensure they can be developed safely. We refer to this ability as an *Off Switch*: the ability to enable a global, coordinated halt on AI. There is not currently consensus about the need to halt AI development, but we believe that it should be uncontroversial that the U.S. should have the option to halt, if and when it decides to do so. Additionally, Off Switch infrastructure would be useful for aims other than halting frontier AI development; it could help with managing a range of risks from dual-use AI systems by providing the ability to flexibly impose restrictions on AI development, deployment, and proliferation. In this comment, we outline R&D directions that would be useful for developing the Off Switch.

For brevity, we aim to limit this response to important R&D directions well suited to government funding. We have previously written a more detailed discussion of the risks from advanced AI and how we believe they should be avoided in our RFI submission on the Development of an Artificial Intelligence (AI) Action Plan and our AI governance research agenda.

Tracking and Measuring Risks from AI Systems

The Federal government has an important role in facilitating research into understanding and mitigating risks from increasingly capable AI systems, including those that could pose significant national security or societal threats.

- **Consensus Tripwire Capabilities**: Invest in R&D to define and operationalize clear, measurable "tripwires" which indicate dangerous AI capabilities. This includes:
    - Capabilities related to WMD (Chemical, Biological, Radiological, Nuclear) development or cyber offense.
    - Demonstrable early misalignment or loss-of-control indicators in advanced models. This could include research into "honeypot" systems and testing environments.
- **Privacy-Preserving Inference Monitoring**: Advanced AI systems may be able to be misused, potentially causing catastrophic harm. Some of this risk may be mitigated by monitoring the inputs and outputs of AI systems. Research is needed for robust privacy-preserving monitoring, which would help prevent misuse while enabling the beneficial use of AI systems.
- **Criteria for Restricting Open Release of Model Weights**: Develop evaluation criteria to guide decisions on when the open release of powerful AI models poses unacceptable risks. This research should inform policies that balance the benefits of openness with national security and public safety concerns. We do not believe that any currently openly released models threaten these concerns, but future models may. AI developer Anthropic recently developed a model, Claude 4 Opus, that reaches a capability level where they state "we cannot clearly rule out ASL-3 risks"; specifically, the model could potentially enable novices to develop a catastrophic biological weapon.

Tracking AI Hardware

Specialized AI hardware (e.g., GPUs, TPUs) is essential for developing and deploying frontier AI. The U.S. government should take an active role in tracking this hardware, domestically and internationally.

- **Track AI Hardware**: The federal government should begin tracking all large stockpiles of AI-relevant hardware, both domestically and internationally.
- **Establish a Comprehensive AI Chip Registry**: Fund R&D into the technical and logistical frameworks for a secure AI chip registry. This registry should aim to track and log the production, distribution, and location of significant quantities of advanced AI chips, both domestically and globally. This would enhance visibility into compute concentrations, a key strategic resource for future AI development.

Technical Mechanisms for Verification

To enable trust, accountability, and the ability to provide the assurances required by potential future international agreements regarding AI development, the U.S. should lead R&D into robust verification technologies.

- **Flexible Hardware-Enabled Guarantees (FlexHEGs)**: Prioritize research and prototyping of FlexHEGs and similar hardware-based governance mechanisms. This includes:
    - Location Attestation: Securely verifying the physical location of AI hardware.
    - Code Attestation: Verifying the integrity and authorized nature of software/models running on hardware.
    - FLOP Counting/Compute Usage Monitoring: Securely measuring the amount of computation used, potentially to enforce training limits or detect unauthorized activity.
    - Tamper-Evident and Tamper-Proof Secure Enclosures: Developing physical security for AI chips that can host governance mechanisms.
    - Retrofittable FlexHEGs: Researching methods to apply governance mechanisms to existing hardware infrastructure where feasible.
- **AI-Enabled and Privacy-Preserving Auditing Technology**: Develop privacy-preserving methods for auditing AI developers and performing AI evaluations. This could include using AI systems to monitor internal AI development. These automated auditors could catch prohibited or dangerous activities, while preserving privacy.

Government Preparedness

The federal government should understand the current AI development landscape and must be prepared to respond to AI-related emergencies.

- **Visibility and Reporting from AI Developers**: Support R&D into standardized, secure, and meaningful reporting mechanisms for AI developers, particularly those working on frontier models. This ensures appropriate government bodies have situational awareness of rapid capability advancements or emerging risks.
- **Emergency Response Planning for AI-Enabled Disasters**: Fund scenario-based research and planning for AI-related emergencies. This includes:
    - Developing mechanisms for tracking and identifying rogue autonomous AI systems.
    - Researching and establishing protocols and technical capabilities for rapidly and safely shutting down or isolating compute resources, both at the datacenter level and, if necessary, for more distributed systems.

Security

Security to prevent the leakage or theft of important AI assets (such as model weights, code, and algorithmic insights) is important to prevent the proliferation of dangerous capabilities.

- **Security of AI Model Weights and Algorithmic Secrets**: Fund R&D into advanced cybersecurity measures, information compartmentalization strategies and other security specifically designed to protect highly capable AI model weights and critical algorithmic insights from theft, unauthorized access, or exfiltration by adversaries or models themselves.

Secure AI Research Facilities

Secure AI research facilities may be useful for beneficial AI research (such as alignment research) while helping to mitigate some of the risks from these systems. Differential access practices may be used to allow some researchers to access advanced AI models, without making these models widely available.

- **Differential Access via the National AI Research Resource (NAIRR)**: Invest in making the NAIRR a testbed for differential access models. This would involve providing tiered, secure access to frontier models and compute resources for vetted researchers (e.g., safety researchers, national security analysts, academics working on beneficial applications) before broader release, enabling pre-deployment testing and related research.

A more detailed discussion of the risks from advanced AI and how we believe they should be avoided, can be found in our RFI submission on the Development of an Artificial Intelligence (AI) Action Plan and our AI governance research agenda.