PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mba-9eiw-aj6n
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-0347
Comment on FR Doc # 2025-07332

Submitter Information

Organization: Institute for Progress (IFP)

General Comment

See attached file(s)

Attachments

IFP - AI Research and Development RFI Response

1255 Union St NE, Suite 820, Washington, DC 20002
ifp.org
contact@ifp.org

Catalyzing a Golden Age: A Blueprint for Strategic AI R&D Investment

Response to the OSTP RFI on the Development of a 2025 National AI R&D Strategic Plan

May 29, 2025

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

About IFP

The Institute for Progress (IFP) is a non-partisan think tank focused on innovation policy. Our organization works to accelerate and shape the direction of scientific, technological, and industrial progress. Headquartered in Washington, DC, IFP works with policymakers across the political spectrum to make it easier to build the future in the United States.

Introduction

AI has the potential to solve some of humanity's most pressing problems, from finding treatments for crippling diseases through accelerated drug discovery, to eliminating food insecurity by engineering cheap and plentiful foods, to delivering a new scientific revolution through the discovery of new materials and tools. The artificial intelligence of the future could catalyze a new golden age of growth, prosperity, and abundance. But to effectively harness these capabilities, we must solve two broad problems.

First, these benefits may not come by default or quickly enough, given existing commercial incentives. This may be because a particular application would create public goods, which markets tend to undersupply, or because a research direction is high-risk and requires large upfront investments. For example, markets alone may not create an AI to automate the replication of scientific studies, or to investigate new medical treatments that can't be patented, or invest in basic research in neuroscience with no immediate commercial applications.

Second, rapidly improving AI capabilities will likely come with risks that industry isn't sufficiently incentivized to solve. AI systems that can broadly accelerate the pace of medical research could also help engineer biological weapons. Advanced coding agents used throughout the economy to vastly increase productivity could also be put to work, day and night, to find and exploit security vulnerabilities in critical infrastructure. Leading AI labs have some incentives to prevent the misuse of their models, but the offense-defense balance of emerging AI capabilities in areas like cyber and bio is uncertain — private incentives to adequately invest in preventing misuse could be dwarfed by the scale of the risks new AI technologies could impose on the public.

The default path of AI development may not deliver us these benefits, nor protect us from these new threats. Taking the default path means surrendering our future to the shape of the AI technology tree: whatever is easiest and most profitable to build will be built first, with no guarantees that this will lead to human flourishing.

But there is an alternative: the US government, as the R&D lab of the world, has long shaped the direction of technological progress, from investing in clean energy to catalyze the massive advances in solar and wind power we see today, to laying the groundwork for the internet, GPS, and modern computing through decades of federally funded research.

The US government can lead the way to a golden age of AI discovery by promoting the development of technologies that bring us scientific benefits faster, and enhancing technologies for safety and security ahead of the development of potentially destabilizing technologies. This will mean developing new public goods that are only imaginable in a world with advanced AI; investing in AI-accelerated vaccine discovery before AI systems can help engineer biological weapons; and investing in automated code refactoring to harden our critical cyber infrastructure before superhuman AI hackers are developed.

Below, we outline key areas where we expect targeted federal R&D investments to yield exceptionally high returns, either by **developing new public goods** that would not be properly incentivized by market forces alone or by **accelerating the development of technologies that would increase our civilizational resilience to new threats**.

Part I: Using AI to accelerate science and build public goods

1. X-Labs: Institutional block grants to maximize the potential of AI for science

See Launching X-Labs for Transformative Science Funding for more information.

Some of the most groundbreaking AI for science initiatives today are not emerging from traditional academic laboratories but from independent research institutions. DeepMind's AlphaFold revolutionized protein structure prediction, and the Arc Institute's Evo 2 enabled

breakthrough genomic analysis. These successes share a common foundation: organizations structured specifically for ambitious, team-based research with stable funding, specialized computational resources, and freedom from traditional academic constraints.

To replicate these breakthroughs at scale, we must address a structural mismatch in our research funding ecosystem. The majority of federal science funding flows through mechanisms designed for individual investigators pursuing discrete, short-term projects — a model fundamentally ill-suited for transformative AI initiatives.

The NIH R01 grant, the workhorse of biomedical research funding, exemplifies this mismatch. R01s typically fund individual principal investigators to work on a specific, pre-specified project for 3-5 years with budgets of around $600k in total over that period. This structure creates multiple barriers to AI breakthroughs:

- Resource Constraints: Training foundational AI models requires computational infrastructure and datasets costing far more than entire R01 budgets. AlphaFold's development required years of sustained investment that no traditional project-based grant could support.
- Expertise Fragmentation: AI for science requires teams that span machine learning, domain science, software engineering, and data infrastructure — expertise that rarely exists within single university labs funded through project-based grants.
- Structural Limitations: R01s weren't intended to fund the shared computational infrastructure, curated datasets, and specialized platforms that AI initiatives require, leaving researchers to cobble together inadequate resources through overhead or philanthropic support.

Similar constraints affect NSF's portfolio. And while NIH "P-Series" and NSF Science and Technology Centers nominally support larger efforts, in practice, they function as collections of individual projects stapled together rather than unified institutions with strong vision and leadership.

We propose the "X-Labs" initiative as a vital framework to address these structural limitations and catalyze impactful AI for science initiatives. X-Labs would fund independent research institutions, selected competitively and awarded via long-term, flexible block grants (between $10M and $50M per year) initially reallocated from existing science agency budgets. This model is designed to accelerate team-based, high-risk, high-reward science in areas like AI that require intensive infrastructure and tooling. Implementation could begin immediately without requiring congressional action by utilizing Other Transactions Authority (OTA) at NSF, NIH, and DOE, starting with ~1% of the science budgets as a pilot.

This initiative would:

- Help seed new scientific institutions specifically built for AI from the ground up, creating research organizations designed around the computational, data, and talent requirements of modern AI development
- Give scientific institutions the flexibility on spending around infrastructure and personnel they need, enabling investments in specialized computing clusters, large-scale data acquisition, and competitive compensation for AI talent that traditional grants rarely support
- Sidestep traditional university bureaucracy, allowing faster execution, reduced administrative overhead, and more agile research planning essential for keeping pace in the rapidly evolving AI landscape

The X-Labs Initiative would establish four complementary funding categories tailored to different aspects of the AI R&D ecosystem:

- **X01 (Excellence) Awards** would support cutting-edge basic science institutions with flexible research environments modeled after organizations like the Janelia Research Campus, the Arc Institute, and the Broad Institute. These institutions would focus on foundational scientific discovery with stable, long-term support. The core bet behind X01s is on a group of people, not projects — the goal is to assemble the best *team* in the world to pursue open-ended scientific inquiry with minimal bureaucratic constraint.
- **X02 (Execution) Awards** would fund focused research organizations oriented toward solving critical infrastructure, tooling, or data challenges. This would involve funding a talented group with a nimble organizational structure to execute against a clearly defined bottleneck in the scientific ecosystem.
- **X03 (Experimentation) Awards** would support portfolio-based funding organizations that can scout and support promising research directions more nimbly than traditional government grant processes. These awards would empower *scientific scouts:* Individuals or organizations with the insight, network, and conviction to identify high-potential ideas, talent, or research directions long before they become consensus picks.
- **X04 (Exploration) Awards** would provide $1–3 million in seed funding to support the formation of new scientific institutions, enabling teams to refine their vision, build key partnerships, and develop initial proof-of-concept work before applying for full X01, X02, or X03 funding.

X-Labs would complement traditional project-based funding by creating an institutional layer in the federal research portfolio specifically designed for the team-based, infrastructure-intensive nature of modern AI development. This approach would ensure that America's most talented scientists and researchers can focus on breakthrough AI work rather than navigating grant cycles and administrative requirements. If we want to be able

to take full advantage of the new opportunities presented by advanced AI, our institutional and funding mechanisms will have to evolve alongside it.

2. A million-peptide database for AI-enabled antibiotic discovery

See How Scientific Incentives Stalled the Fight Against Antibiotic Resistance, and How We Can Fix It.

For nearly all of human history, infectious disease has been our deadliest foe. In the first decades of the 20th century, nearly one in a hundred Americans would die of an infectious disease every year.

In recent decades, this number is about one hundred times lower (around one in ten thousand). This is in part thanks to antibiotics — medicines used to fight infections. However, the growing problem of antibiotic resistance — driven by a combination of increasing use of antibiotics and a lack of new antibiotics under development — threatens a return to pre-antibiotic mortality rates. Antibiotic resistance already kills over 1.2 million people annually worldwide.

Antimicrobial peptides are especially promising candidates to deal with this crisis. They're naturally resistant to bacterial evolution, easily "programmable" by changing their amino acid chains (which makes them easy to work with through predictive machine learning methods), and can be manufactured in days rather than years. Yet despite their promise, peptides remain absent from pharmacy shelves.

Even though the molecular chains of peptides are short (usually less than 50 amino acids), the combinatorial space of peptide sequences is vast. It's difficult to search this space for peptides that effectively combat the antibiotic-resistant "superbugs" that threaten to massively increase the rate of deadly infections. This problem is well-suited to machine learning methods, which could be used to predict the structure of new antimicrobial peptides. The main constraint is gathering enough data. While DeepMind's AlphaFold revolutionized protein structure prediction by training on 100,000+ 3D protein structures in the Protein Data Bank, researchers working on antimicrobial peptides have access to only a few thousand experimentally validated sequences scattered across poorly maintained databases. Current ML approaches show promise, with models already predicting peptides active against MRSA and HIV, but these models are starved for the data they need to unlock breakthrough treatments.

A federal investment of $350 million over five years could create a million-peptide database, amounting to a 1,000x increase over existing resources. Using high-throughput synthesis methods like SPOT, which can screen thousands of peptides at less than 1% of the cost per peptide, a single automated setup can synthesize and test 8,000 peptides every two weeks. This isn't a moonshot requiring breakthrough technology; it's infrastructure that can be built today.

Following the model of the Human Genome Project and the Protein Structure Initiative (which enabled AlphaFold), this database would catalyze an ML renaissance in antimicrobial research. Just as PubChem's 118 million compounds database transformed computational chemistry, a million validated peptide sequences would provide the foundation for AI systems to design treatments for antibiotic-resistant infections — potentially saving millions of lives while costing less than the annual treatment burden of just six drug-resistant antimicrobial resistance threats in the US.

3. Other AI for science initiatives

The Institute for Progress is collaborating with leading experts to develop concrete yet ambitious R&D "moonshot" proposals that leverage advanced AI to accelerate scientific progress. These will be published on our website (ifp.org) in July, 2025. Below is a preview of these forthcoming proposals:

- **"Automating Scientific Replication Studies":** How to create the infrastructure to have specialized AI agents test the robustness of research findings when they're submitted for review, by Abel Brodeur and Bruno Barbarioli.
- **"Scaling materials discovery with self-driving labs"**: How to close the gap between AI-guided material design and real-world validation by building automated experimental platforms, by Charles Yang.
- **"A Stargate Project for Biology":** How to use AI for hypothesis generation and support hundreds of fast clinical trials to test treatments to rid humanity of 99% of its disease burden, by Sam Rodriques.
- **"Enabling At-Scale Pathogen Surveillance with AI":** How to generate the data necessary to reliably detect new pathogen outbreaks with AI, by Simon Grimm.
- **"Connectomics for AI":** How to map out the mammalian brain's connectome to solve fundamental problems in neuroscience, psychology, and AI alignment, by Adam Marblestone.
- **"Teaching AI How Science Actually Works":** How to create for physical science what the internet is for the digital world by recording real research at scale, by Ben Reinhardt.
- **"A National Evaluation and Challenge Institute":** How to create TELOS, a pilot institution, to systematically commission AI evaluations and grand challenges, leveraging a demand-pull mechanism to steer global AI research toward US leadership, basic science, and the public good, by Séb Krier and Zhengdong Wang.

Part II: Making AI secure and reliable

1. A "Human Genome Project" for AI interpretability

AI interpretability research aims to develop a more concrete understanding of a model's predictions, decisions, or behavior. Solving interpretability will allow for safer and more effective AI systems via more precise control, and help in detecting and neutralizing adversarial modifications such as hidden backdoors. Interpretability techniques can also be used to extract novel scientific insights from neural networks that traditional analysis methods cannot discover – for example, the Arc Institute is using interpretability techniques on their Evo 2 model to reveal novel biological mechanisms and patterns that the model learned implicitly from training data.

Early interpretability research suggests we may be on the cusp of meaningful theoretical breakthroughs. But the scale and urgency of this challenge demand a more ambitious approach than existing grant programs. A large-scale initiative — comparable in ambition to the Human Genome Project — could be instrumental in solving interpretability.

Given the strong overlap of this work with defense interests (including increasing the reliability of AI models deployed in national security applications, and understanding the capabilities of adversary systems), this work could be coordinated through defense agencies and spending, using target product profiles from the defense and intelligence communities that set clear parameters on the kinds of interpretability features they would like from an AI model or application. A "grand challenge" to develop new solutions could then be supported through proven, efficient funding mechanisms, such as:

- Prize competitions for novel interpretability research techniques, with tiered prizes for different aspects of interpretability (e.g., circuit discovery, concept visualization, neural network decomposition), and
- Challenge-based acquisition programs and advance market commitments (AMCs), involving commitments to purchase technical solutions that successfully meet certain criteria

There isn't enough consensus in the field of interpretability as to which methods will deliver the desired results. AMCs and prize competitions have the advantage of not centralizing R&D around a single methodology or research organization, instead resulting in a decentralized R&D effort where the best methods win — and the government doesn't need to spend any money unless its goals are met.

2. Hardware-backed security and verification

AI chips are crucial to the development of advanced AI and the ability to diffuse AI throughout the economy. The US is the undisputed leader in the AI chip market, producing almost all of the world's most advanced AI chips.

Since October 2022, the US has sought to control access to these chips with export controls. While these controls have been broadly effective at preventing China from amassing large quantities of AI compute, extensive smuggling shows that they are also being circumvented. While more stringent export controls applied to a growing number of countries could help prevent smuggling, this would also create large burdens for the US semiconductor industry.

Today's hardware and software technologies, such as Confidential Computing, can help strike a better balance of security and competitiveness. For example, software-based location verification could enable the US to continue to export advanced AI chips to allied and neutral countries while cracking down on smuggling into the People's Republic of China (PRC). Privacy-preserving workload verification could help computing providers abroad (such as the UAE's G42, which is now set to build some of the largest and most advanced AI data centers in the world) prove that their chips aren't being used to train highly advanced models for PRC developers. These approaches complement the Trump Administration's priorities of tough export controls that don't disadvantage US industry and that enable the diffusion of US technology infrastructure.

These technologies can also protect American AI infrastructure (including chips, data centers, and models) from industrial espionage and sabotage. So far, the development of advanced AI chips has focused on making them as performant as possible, rather than prioritizing security. While the latest AI chips from NVIDIA include some important security and privacy features, like Confidential Computing, these are not yet robust enough to secure IP at the whole-cluster level, or to defend against physical attacks to steal sensitive data. And while NVIDIA and other companies are heavily incentivized to produce the most advanced AI chips, they are not strongly incentivized to make sure that the chips are harder to smuggle into China, or that chips include the necessary hardware security features to securely verify compliance with export controls.

The US government is well-positioned to bridge this gap. Programs such as the National Semiconductor Technology Center (NSTC), the Department of Defense's Microelectronics Commons, DARPA's Microsystems Technology Office (currently pursuing multiple relevant projects), and NIST's long-standing leadership in hardware security standards can serve as focal points for accelerating research and implementation. DARPA and/or the Commons should run a series of challenge prizes, as a public-private initiative to attract co-funding from industry, to develop:

- Cluster-level trusted execution environments to enhance user privacy, protect AI model weights, and create a platform to enable zero-knowledge attestation of AI workloads. This technology could also be used to verify compliance with export controls and other end-use policies (e.g., whether a cluster being remotely accessed by a PRC developer is being used to train a frontier model) in a privacy-preserving way.
- Security modules on AI chips and chip/server enclosures that are robust against adversaries with physical access to the chips, including tamper resistance and protection against side channel attacks.
- Delay-based location verification to detect and deter AI chip smuggling.

See the appendix in *Technology To Secure the AI Chip Supply Chain* for more details on this hardware security and verification agenda.

3. A government research cluster as a test range for AI security

As AI systems become more central to scientific progress, economic growth, defense, and intelligence, the security of the data centers that house these systems must be treated as a national priority. Today's AI data centers are not prepared to defend against well-resourced adversaries, making it possible for them to steal AI model weights. Successfully defending models will be critical for American AI leadership if the predictions of many AI researchers are correct, and AI models themselves become the primary driver of AI progress by automating the process of AI R&D. AI data centers must also be made more resilient to denial or sabotage operations — as AI systems are increasingly integrated into the economy and critical infrastructure, data centers will likely become increasingly attractive to attackers as a point of vulnerability for offensive operations against the United States.

Securing AI data centers presents a somewhat different challenge than securing conventional computing infrastructure. Existing high-security data centers, such as those used for classified government operations, prioritize confidentiality and controlled access, but do not have strong performance and scale requirements. Advanced AI data centers operate at a different scale, with specialized infrastructure — including high power-density GPU servers, high-bandwidth networking, and unique cooling requirements — that is optimized for performance rather than security. This creates a security gap that must be addressed.

A 2024 RAND report on AI security laid out a framework for model weight protection, with "Security Level 4" (SL-4) defined as the threshold at which it is possible to defend against routine attacks from top-tier cyber adversaries. This level of security does not currently exist in practice at a single AI training data center. An SL-4 data center is likely achievable within the next few years, but reaching it will require targeted investments in secure architectures, access controls, and best practices for AI model protection. Foreign nationals

have already stolen trade secrets from leading AI labs, and these technologies will become even more valuable targets as AI capabilities grow.

The DOD should build and operate a research cluster to develop best practices for securing sensitive AI workloads and models, providing a "test range" for startups and academic researchers to test new research ideas and technologies for AI security across the software and hardware stack. This facility would serve as a testbed for next-generation security measures, including advanced access controls, red-teaming protocols, and infrastructure monitoring. The overall goals for the cluster should be to conduct research and develop technical standards useful for reaching SL-4, and solving trade-offs between security and performance.

3. Other AI security and reliability initiatives

As part of our "AI Moonshots" project mentioned above, IFP is also collaborating with leading experts to develop concrete yet ambitious R&D "moonshot" proposals that could dramatically improve AI security and reliability. These will be published on our website (ifp.org) in July, 2025.Below is a preview of these forthcoming proposals:

- **"Preventing the Deployment of AI Sleeper Agents":** How to ensure American AI models are robust and reliable by setting up a large DOD-led red- and blue-teaming effort, by Evan Miyazono.
- **"The Great Refactor: Automated Code Hardening at Scale":** How to secure critical open-source code against memory safety exploits by automating code refactoring with AI, by Herbie Bradley and Girish Sastry.
- **"Automated Cybersecurity"**: How to use frontier AI models to find and patch code vulnerabilities at scale before the diffusion of these AI capabilities, by Miles Brundage.
- **"Redesigning the AI Hardware Stack for AI Security":** How to redesign the AI hardware stack to enable better compute policy options and privacy-preserving verification, by Nora Ammann and David "Davidad" Dalrymple.
- **"Building a Scalable, Secure AI Data Center":** How to build a highly secure AI cluster to protect AI model weights, by Sella Nevo.