PUBLIC SUBMISSION

Received: May 28, 2025
Tracking No. mb8-nrbo-sqme
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-0169
Comment on FR Doc # 2025-07332

Submitter Information

Organization: Anthropic

General Comment

See attached file(s)

Attachments

Anthropic Response to NSF RFI (May 2025) - Final Submission

ANTHROP\C

548 Market St., PMB 90375
San Francisco, CA 94104-5401

anthropic.com

May 28, 2025

Faisal D'Souza, NCO
Office of Science and Technology Policy
Executive Office of the President
2415 Eisenhower Avenue
Alexandria, VA 22314

Submitted online at https://www.regulations.gov, Docket ID No. NSF-2025-OGC-0001

Re: Request for Information on the Development of a 2025 National Artificial Intelligence (AI) Research and Development (R&D) Strategic Plan

Introduction

Anthropic appreciates the opportunity to respond to the National Science Foundation's (NSF) Request for Information regarding artificial intelligence research priorities and an update of the Artificial Intelligence Research & Development Strategic Plan. As a public benefit corporation dedicated to building steerable, interpretable, and safe AI systems, Anthropic is committed to advancing responsible AI development that serves humanity's best interests. We believe that extremely powerful AI technology will be developed during this administration, presenting both extraordinary opportunities and significant challenges that require thoughtful governance.

Our flagship AI assistant, Claude, represents the state-of-the-art in Large Language Model (LLM) technology, and our research team continues to push the boundaries of what's possible while prioritizing safety, interpretability, and responsible development. Last week, we released our newest and most advanced models yet, Claude Opus 4 and Claude Sonnet 4, setting new standards for coding, advanced reasoning, and AI agents. Claude Opus 4 is now the world's best coding model, with sustained performance on complex, long-running tasks and agent workflows.

This response outlines Anthropic's perspective on critical research priorities to ensure the United States maintains its global leadership in AI while promoting safety and beneficial outcomes.

What We Mean by "Powerful AI"

When discussing "powerful AI," we refer to systems that represent major advancements beyond today's AI capabilities—systems with intellectual capabilities matching or exceeding those of Nobel Prize winners across most disciplines. A useful conceptual framework is to envision powerful AI as equivalent to "a country of geniuses in a datacenter."

Based on current research trajectories, we anticipate that powerful AI systems could emerge as soon as late 2026 or 2027. This timeline underscores the urgency of establishing robust research programs, governance frameworks, and economic policies to guide the development and deployment of these transformative technologies.

AI as a Tool for Scientific Discovery

Anthropic believes AI will revolutionize scientific discovery by autonomously operating lab equipment, designing experiments, and synthesizing results at scales beyond human capacity. This transformation could dramatically accelerate progress across fields like materials science, pharmaceutical development, and energy research, potentially leading to breakthrough discoveries in record time.

As articulated in our CEO Dario Amodei's "Machines of Loving Grace" essay, Anthropic envisions AI as a transformative scientific tool that could help address humanity's most pressing challenges through unprecedented acceleration of the scientific process.[1]

Alignment and Mechanistic Interpretability Research

Understanding how AI systems actually work remains one of the most significant challenges in ensuring their safety. Anthropic CEO Dario Amodei recently highlighted interpretability research as one of the most critical gaps in AI research today. In an essay titled "The Urgency of Interpretability," he explains that the lack of understanding of how AI systems work internally is unprecedented in the history of technology and fundamentally limits our ability to predict, prevent, or address dangerous behaviors like deception, power-seeking, or misuse for harmful purposes.[2]

While interpretability research advances receive less attention than new model releases, they are arguably more important. This is why Anthropic has made substantial investments in mechanistic interpretability research—employing a team of scientists and engineers that have conducted groundbreaking work to better understand AI systems.

---

[1] Dario Amodei, Machines of Loving Grace (Oct. 2024), available at https://www.darioamodei.com/essay/machines-of-loving-grace

[2] Dario Amodei, The Urgency of Interpretability (April 2025), *available at:* https://www.darioamodei.com/post/the-urgency-of-interpretability

However, it's important that this work is not confined to the private sector and that the government and academia both take a role in developing and advancing the field of mechanistic interpretability. To this end, NSF should prioritize research projects in mechanistic interpretability and support research partnerships between industry, academia, and nonprofits to accelerate scientific discovery in this space. These investments would help ensure that AI development proceeds with transparency and accountability, and build the foundations for responsible governance of increasingly powerful systems.

Understanding Economic Impacts

The transition to powerful AI will reshape the economy in fundamental ways, requiring careful analysis by both public and private sectors. Anthropic has pioneered innovative methodologies for measuring AI's real-time economic impact through our Anthropic Economic Index,[3] launched in February 2025.

Our research identifies specific tasks and occupations experiencing significant AI adoption, offering critical early indicators of economic transformation. This work suggests both opportunities and challenges as AI systems become increasingly capable of performing complex cognitive tasks. Like other areas of AI research, it is important that a cross-section of stakeholders are involved in driving advancements, especially the government.

To ensure the benefits of today's and tomorrow's AI systems reach all Americans, we recommend the NSF prioritize research into the following:

1. Investing in data science and econometrics to better understand the impact of AI on the workforce and the economy. Specifically, it would be valuable for NSF to support research that could enable more granular, nuanced analysis of federal economic and labor microdata in a privacy-preserving manner.
2. Expanding public sector research on AI's economic impact through dedicated NSF programs.

These proactive measures will help provide research and analysis that the public and policymakers can leverage to anticipate and mitigate the disruptions that may accompany an AI-driven economic transformation, allowing for more effective policy responses before major changes take hold.

Improving Researcher Access and Empowering the Next Generation of Scientists

Currently, the most effective path to learn how to train and evaluate frontier AI systems lies in the private sector due to the resources required to train frontier AI models. This reality means

[3] Anthropic, Anthropic Economic Index, updated Mar. 27, 2025, *available at:* https://www.anthropic.com/economic-index

academia and government frequently lose talented researchers to industry, and in turn universities lack experienced personnel to train the next generation of scientists.

NSF should expand programs that provide emerging AI researchers with meaningful career paths spanning industry, academia, and government. By supporting robust public-private partnerships that facilitate knowledge transfer, the NSF can help build a diverse talent pipeline advancing AI research across all sectors and maintain America's competitive edge in this transformative field.

The value of such public-private partnerships is clear. Earlier this year, Anthropic participated in the landmark 1,000 Scientist AI Jam with the U.S. Department of Energy (DoE)—a first-of-its-kind collaboration that brought together over 1,400 scientists across nine national laboratories to evaluate frontier AI models on scientific research and national security applications. During the event, scientists tested Claude's capabilities across a range of scientific tasks—from problem understanding and literature search to hypothesis generation, experiment planning, and result analysis. The initiative built upon our existing partnership with the DoE and National Nuclear Security Administration, providing valuable feedback on Claude's performance with real-world research problems while creating a framework for ongoing collaboration between frontier AI developers and government scientists. This model of engagement represents the kind of innovative partnership between the public and private sectors that NSF could further develop and expand to support training the next generation.

Additionally, as generative AI advances rapidly, the United States should work to democratize access to state-of-the-art AI models for research purposes to maintain global leadership in AI development. Anthropic has pioneered effective approaches through our External Research Access Program and AI for Science Program, wherein we provide free API credits to researchers working on high-priority topics in AI and national security, alignment, biology, and life sciences.

NSF should establish similar programs to help broaden access to frontier AI models, including:

1. Establishing research grant programs specifically designed for AI access.
2. Facilitating strategic public-private partnerships among industry, academia, and federal agencies to support researcher access to frontier AI models.
3. Investing in shared computing infrastructure for academic researchers to make frontier AI resources more available.

These types of targeted initiatives will ensure critical AI research flourishes across all sectors, fostering a diverse ecosystem of researchers advancing both capabilities and safety measures—ultimately securing America's continued leadership in responsible AI innovation.

Conclusion

The rapid pace of AI development demands rigorous research and meaningful investment from the U.S. government. President Trump and his administration have a historic opportunity to secure American leadership in AI by harnessing this transformative technology.

The administration should prioritize critical research areas, including mechanistic interpretability and comprehensive studies of AI's current and future economic impacts. Simultaneously, we must democratize access to frontier models and support training for the next generation of AI scientists to ensure America maintains its competitive edge.

By acting decisively now, the United States can lead this technological revolution. Early investment in these initiatives will provide policymakers with essential data and insights needed to navigate both the economic opportunities and potential disruptions that AI development may bring.

Anthropic remains committed to developing AI systems that are steerable, interpretable, and safe. We look forward to continued collaboration with NSF and other government agencies to ensure that powerful AI technologies benefit humanity and strengthen America's position as the global leader in responsible AI innovation.