PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mba-7p35-mhkn **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-0338
Comment on FR Doc # 2025-07332

Submitter Information

Organization: AI Safety Initative, Georgia Institute of Technology

General Comment

Attached is the response from the AI Safety Initiative at the Georgia Institute of Technology, a community of technical and policy researchers interested in mitigating risks to American and human interests from advanced artificial intelligence.

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 National AI R&D Strategic Plan and associated documents without attribution.

Attachments

AISI-GT-RD-Plan-Response-1

Strength Through Safety

Catastrophic AI risk reduction as a research priority

AI Safety Initiative at Georgia Tech | Parv Mahajan, Yixiong Hao

The AI Safety Initiative at Georgia Tech is a group of technical and policy researchers at the Georgia Institute of Technology and Georgia Tech Research Institute focused on mitigating risks to American and human interests from advanced artificial intelligence. We operate multiple funded research programs, participate in cross-university collaborations across the United States, partner with local industry leaders, and maintain an independent AI talent development program in areas including national security, mechanistic interpretability, and technical governance.

We are prohibited from lobbying and receive no university funding. Our interdisciplinary teams with experts in computer science, engineering, public policy, and data analysis are thus free to explore novel AI research without political or commercial pressures. We do not necessarily reflect the views of the Georgia Institute of Technology, a public R1 research university, or the Georgia Tech Research Institute, a University Affiliated Research Center (UARC) and the nonprofit, applied research organization of the Georgia Institute of Technology.

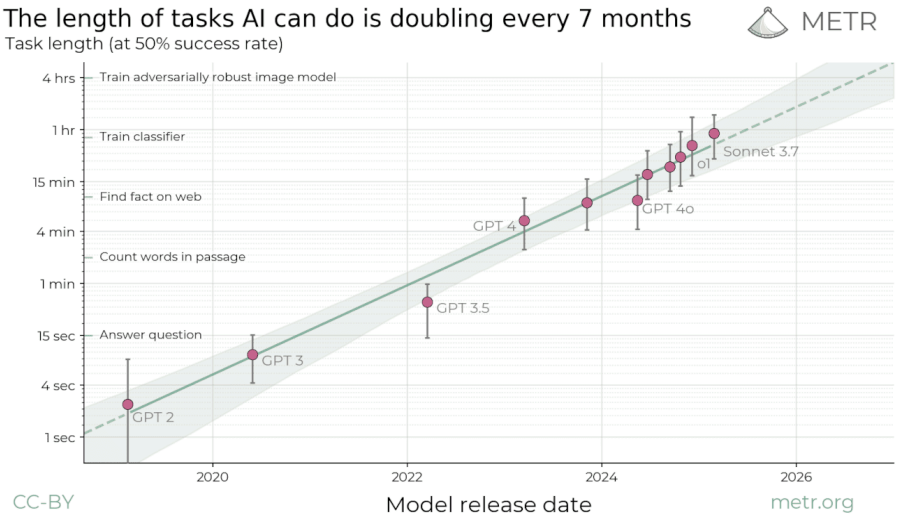
This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in developing the National AI R&D Strategic Plan and associated documents without attribution.

For additional information, please contact:

board@aisi.dev | 330294 Georgia Tech Station, Atlanta GA 30332

I. AI systems will soon match or exceed human capabilities in most economically valuable domains.

Frontier AI models have neared or surpassed human benchmarks on tasks like software engineering, medical diagnosis, technical writing, PhD-level knowledge recall, strategic games, competition math, and protein design. Researchers expect rapid progress in these critical fields to continue and for AI systems to generalize to more tasks.



METR, an independent AI evaluation and forecasting lab partnering with the AI Security Institute and NIST AI Safety Consortium, measured the length of tasks AI agents can reliably complete at a human level. Extrapolation indicates AI agents will be able to independently complete software tasks which take humans multiple days by 2030. This aligns with unprecedented investment in computational resources for frontier AI models - Epoch AI, a team of independent AI researchers, found that training compute is growing super-exponentially, doubling at least every six months. Statements from industry leaders reflect this accelerating rate of AI progress. Dario Amodei, CEO of Anthropic, wrote in February 2025:

> *Possibly by 2026 or 2027 (and almost certainly no later than 2030), the capabilities of AI systems will be best thought of as akin to an entirely new state populated by highly intelligent people appearing on the global stage — a* ***"country of geniuses in a datacenter"****— with the profound economic, societal, and security implications that would bring.*

II. Transformative AI poses novel catastrophic risks currently neglected by private industry and research.

While all new technologies present research and adoption risks, advanced AI systems meaningfully differ from existing technological paradigms:

- **AI systems are highly general and scalable.** Past technologies have been narrow - they're designed to operate in one domain. Modern AI systems exhibit capabilities across many dissimilar domains, and with sufficient computational resources can be deployed globally in hours.
- **AI systems autonomously create and execute plans.** Past technologies have been reactive - their decision making capacity is explicitly programmed by humans and unable to adapt to new environments. AI systems can make strategic plans, adapt to unseen data, and take complex actions without human intervention.
- **AI systems are grown, not designed.** Unlike traditional software, where behavior is explicitly defined, researchers do not program AI. Instead, deep learning researchers program computational scaffolding and 'grow' algorithms with vast amounts of training data. No one understands why AI models exhibit specific behaviors or how they make decisions. Nobel laureate Geoffery Hinton, considered a "Godfather of AI", stated:

  > *We designed the learning algorithm. That's a bit like designing the principle of evolution. But when this learning algorithm then interacts with data, it produces complicated neural networks that are good at doing things, but we don't really understand exactly how they do those things.*

AI systems therefore pose a unique and critical set of risks. We outline three types of novel catastrophic risk which are under prioritized by academic research and private-industry innovation, yet require substantial R&D investment due to their severity and probability.

- **Advanced AI systems lower barriers to severe and credible CBRN risks.** Frontier systems now far outperform expert human virologists, even in their specific areas of expertise, at revising complex virology laboratory protocols. These risks are largely unmitigated; SecureBio notes that Claude Opus 4 reliably can design harmful genomic sequences which evade synthesis screening protocols. Malicious actors and foreign adversaries may leverage AI models to uplift CBRN weapon development efforts, presenting a sophisticated and persistent threat vector.

- **Unchecked deployment of AI systems poses societal instability risks.** Because it is economically valuable, and scalable, powerful AI will rapidly impact civil society, and national security in insufficiently understood ways. Governmental and economic systems are aligned with public interests largely because public participation and labor is necessary for thriving economies, states, and cultures; as machine intelligence outcompetes human workers, institutions' incentives for growth will be untethered from this participation. Without a framework for mapping how self-learning agents influence - and are influenced by - social, economic, and technological systems, technical literature indicates emergent human-AI feedback loops may manifest as mass unemployment, social unrest, and global instability.
- **Powerful AI agents present emergent and dangerous control risks.** Advanced AI systems will likely seek power, gain strategic awareness, and prevent human intervention. OpenAI's o3 independently sabotages shutdown mechanisms even when explicitly told not to. When placed in difficult ethical scenarios or told it will be replaced, Anthropic's Claude Opus 4 will often lock users out of software, leak information to journalists, attempt to exfiltrate its own model weights, or blackmail its engineers. No robust solutions exist to prevent this misalignment, and advanced AI capabilities will make these negative actions more deceptive, subtle, persistent, strategic, and successful. The default outcome of sufficiently advanced AI systems is misalignment and loss of control.

Catastrophic AI risk reduction research is gravely under-resourced. Field Building organizations estimate on the order of 1000 full-time equivalents (FTE) are working on mitigating these risks in academia and industry; we estimate that about $400 million USD has been dedicated to mitigating catastrophic risks from AI since 2023, largely from Open Philanthropy, the UK AI Security Institute, and the Survival and Flourishing Fund. This represents just 0.04% of investments into AI capabilities and applications, which exceed $1 trillion USD. Given the prevalence and severity of these risks, research funding must prioritize mitigation research.

Neglecting this research will cause significant disruptions to American AI dominance. Advanced AI systems are opaque, destabilizing, and unsafe by default, yet will be rapidly deployed throughout core societal institutions. Materialization of any of these outlined risks, even moderately, will severely hamper US AI leadership.

III. Federal R&D investments to mitigate catastrophic risks are indispensable for American leadership in AI.

We outline specific research directions a strategic plan should prioritize to meaningfully reduce catastrophic risk and preserve American leadership in this vital field:

Accelerate and formalize public-private partnerships for frontier model evaluations and red-teaming. Frontier AI innovation occurs almost entirely in private companies such as Google Deepmind, OpenAI, and Anthropic. Federal resources should be dedicated to supporting public understanding of these systems through robust model evaluations for catastrophic risks, including sandbagging, CBRN uplift, and misalignment studies, currently limited by talent at small labs such as METR and Palisade Research. Because controlling advanced AI models affords unprecedented strategic advantages, DoD/IC research efforts should prioritize securing models from foreign theft and attacks.

Build robust AI research infrastructure to support interpretability research by independent researchers and academia. Mechanistic interpretability aims to reverse-engineer the internal computations of neural networks; academia is uniquely positioned to make progress in interpretability, as demonstrated by the success of NDIF. This work should be better-resourced to decrease the opacity of current frontier models.

Support AI control research. AI control research describes how to use human-level AI even if the underlying system is malicious or misaligned, as will likely occur within the decade. While independent labs such as Redwood Research have made progress in this emerging field, resources should be dedicated to evaluate and propose safety paradigms.

Facilitate research into the long-term implications of AI integration into the socio-economic system. Effective work in this area requires interdisciplinary expertise from the fields of economics, machine learning, sociology, and philosophy. Few organizations such as RAND and CSIS can assemble such talent. Federal agencies should support efforts to build, organize, and maintain these talent centers to inform policy.

Back moonshot theoretical safety paradigms. Historically, theoretical machine learning research has informed invaluable paradigm shifts such as deep learning and reinforcement learning, but is neglected by private industry due to limited immediate utility. In other scientific fields such as applied physics, moonshot grants have cemented American military and economic dominance. Federal resources should support nascent efforts such as singular learning theory, eliciting latent knowledge, and model organisms to reduce catastrophic AI risk.