# PUBLIC SUBMISSION

**Received:** May 29, 2025
**Tracking No.** mba-2jn0-1yqg
**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-0305
Comment on FR Doc # 2025-07332

## Submitter Information

**Organization:** ControlAI

## General Comment

Good evening,
Please find comments from ControlAI on Docket ID No. NSF-2025-OGC-0001 Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan attached. We are thankful for the opportunity to comment.

Sincerely,
ControlAI

## Attachments

FINAL ControlAI Comments on Docket ID No. NSF-2025-OGC-0001 Request for Information_ Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

29 May 2025

# Comments on Docket ID No. NSF-2025-OGC-0001 Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

Control AI welcomes this opportunity to comment on the Office of Science and Technology Policy (OSTP) Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan.

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 2025 National AI R&D Strategic Plan and associated documents without attribution

## Who we are

ControlAI is a non-profit and non-partisan organization focused on global security risks from advanced AI systems. Our work to date has focused on policy recommendations for the Bletchley Park AI Summit, advocacy and criticism of the draft EU AI Act, advocacy for policy measures to address the rising impacts of deepfakes, and policy research and advocacy on the risks of building superintelligence, including through our comprehensive policy proposal, "A Narrow Path."[1] We have presented our work to a range of government and nonprofit bodies, and our work has been featured in a wide range of news publications and broadcasts (such as Time Magazine, Bloomberg, GB News, The Daily Mail, and the Guardian)[2]. ControlAI is also a member of the Campaign to Ban Deepfakes, a coalition aiming to reduce growing threats from nonconsensual AI-generated synthetic content.

## Our response

In response to OSTP's RFI, we wish to provide the following comments to inform the 2025 National AI R&D Strategic Plan.

[1] Accessible at www.narrowpath.co
[2] A selection of these media appearances can be found at https://controlai.com/media

## How we see the situation

Our overarching comment is this: there are many current and future types of artificial intelligence that the United States could safely build; these capabilities are compatible with the Executive Order 14179 policy to "sustain and enhance America's global AI dominance in order to promote human flourishing, economic competitiveness, and national security." These varieties, often categorized as "tool" or "narrow" artificial intelligence, would lack the kinds of capabilities that make them dangerous to develop and disloyal to their users, while still being powerful for uses ranging from cancer research to economic growth to national security purposes.

However, building true Artificial General Intelligence (AGI)[3] that can match or exceed humanity at every economically valuable task would severely endanger global and national security. Though AI companies are on a short path to "grow" such AGI, no one knows how to understand it, how to make it loyal, or how to prevent it from being a catastrophic threat to American national security and the survival of the entire human race.

We believe that Artificial General Intelligence is potentially only 1 to 5 years away. This belief is driven not only by a range of outside-in assessments of technical progress and discussions with experts, but also the on the on-the-record statements of frontier AI company CEOs and leadership[4], as well as more private conversations with current and former employees at a variety of levels within those companies.

This is not good news. We must take a different path, one that keeps America in control.

## Our recommendations

### Do not invest in recursive self-improvement

As OSTP's stated intent is to ensure that the 2025 National AI R&D Strategic Plan will put "particular attention on areas that industry is unlikely to address", it is important to understand the *current* investment approach for frontier AI companies.

By the public statements of frontier AI companies,[5] once they have the ability to automate AI research through AGI or near-AGI systems, they intend to hand over substantial responsibility to those AI systems to conduct frontier AI research, with human researchers rapidly losing relevance as big tech companies race to build superintelligent AI. This is despite the fact that, as Anthropic CEO Dario Amodei said in an interview in 2024, we

[3] By AGI, we mean the frequently-used definition of AI capable of doing any intellectual task that a human can do.
[4] For example, though their preferred terminology slightly differs, Anthropic has publicly indicated that their submission to this RFI will also note the potential of very powerful AI being developed within this Administration.
[5] A roundup of several instances of this can be found at https://controlai.news/p/from-intelligence-explosion-to-extinction

understand only about "3%" of how artificial intelligence models work.[6] Once the handover to AGI-driven, automated AI research is made, *no country or company will be in control* of what happens next. Superintelligent AI is too dangerous to its own country to be developed; from the moment it is created, no matter *who* creates it, superintelligence threatens the security and continued existence of the United States.

Building AGI is like building the Doomsday Machine in the movie *Dr. Strangelove*; once you build it, the world is on the edge of disaster at any moment. Rushing ahead and letting a small number of companies in any country build doomsday devices neither enables American competitiveness, nor national security, nor human flourishing – we will all be dead. As with other technologies which pose severe security risks, such as nerve gas or lab-grown plagues, **no private person, company, or government should seek to develop or deploy superintelligent AI systems.**

The US government should not invest *one single dollar* of taxpayer money in enabling the current industry plan to give up control to AI systems that we do not understand. Every grant, loan, contract, and partnership with industry should be screened against this principle.

### Make additional investments to enable chip verification and national security awareness

The US government should begin long-term research and development efforts to identify longer-term needs to maintain and enhance their inspections program via verification mechanisms. At scale, this will require tamper-resistant verification mechanisms throughout the hardware and software stack. Mechanisms could be developed by either the public or private sector alone, but likely will be more robust if they are developed through a process that also incorporates both government security insights and outside input and testing, similar to the US NIST cipher competitions for general-public and government cryptographic use. To be practicable, these mechanisms would need to be capable of reporting signals of dangerous use of large amounts of compute without generally violating the underlying privacy of compute users. For instance, "reporting dashboard enablers"that help track exceptionally large amounts of compute usage by customers over a given threshold would meet this criteria, but backdoors into every processor would not.

### Invest in fundamental insights research needed for maintaining control

The National Artificial Research and Development Strategy should invest in three neglected areas of research in order to be able to keep America in control.

**Prediction of AI systems capabilities:**

The biggest obstacle to safe AI development with current ML technology is the inability to predict what an AI system can and cannot do. This is the case not only before pre-training or fine-tuning, but even after deployment of the AI system. In one example among many, Anthropic testers realized accidentally that their new model was able to recognize it was

[6] Comments when interviewed by Norges Investment Bank, 2024, available at https://www.youtube.com/watch?v=xm6jNMSFT7g&t=161s

undergoing tests and alter its behavior accordingly. And despite extensive efforts to develop theories of Deep Learning, mechanistic interpretability, and evaluation frameworks, still nobody is able to predict what ML models can and cannot do before they are trained. Existing evaluation methods measure, at best, proxies of intelligence only *after* they are developed.

Yet prediction is essential; without this, national security efforts cannot effectively prioritize threats, either from dangerous uncontrollable models themselves *or* from hostile development efforts by adversaries to build models (whether "narrow" tool AI or dangerous, uncontrollable AI).

Taking inspiration from historical examples, the first step to building a science of prediction is to design ways to measure the underlying phenomena. In structural engineering, this came about in the mechanical testing of materials; in aviation, with the measurement of aerodynamic forces, notably in wind tunnels; in nuclear technology, with the measurement of radiation, for example with Geiger counters. In each of these cases, the development of measurement methods was not just about building tools – it also required theoretical and conceptual innovation to figure out what to measure, and how to measure it, to get the right information, often indirectly. Once intelligence can be sensibly measured, the data collected through these measurements will lead to a science of intelligence that can be used for predictive purposes. This will notably include a mechanistic model of intelligence: a decomposition of intelligence into components such that knowing which components are implemented in an AI system lets you predict its intelligence and capabilities in advance before even building it, or turning it on.

**Specification of AI systems capabilities:**

The next step towards building controlled AI systems lies in figuring out exactly which properties they need to satisfy in order to be safe. This might include properties about controlling these systems, about them being legible to users and inspectors, or about them never proposing actions that are particularly unsafe.

Current ML research does not even try to do this, focusing instead on measures of efficiency, performance, and proxies such as "truthfulness." These measures are also constantly being gamed by machine learning systems, since they do not capture specific features of machine learning systems' properties, but merely statistical similarities in large amounts of low-quality data.

We currently do not have a specification language that can write down the things we actually want from AI in detail, nor do we have the ability to model their interactions. A specification language is not enough though: it is also essential to figure out which exact properties we need to express in this language. Since the sole purpose of the specification language is to allow the specification of these guarantees, formalizing these guarantees and designing the language will go hand in hand. In the end, this effort would result in a formal specification language that can address any AI system behavior, including interaction with subcomponents,

other AI systems, and humans. The guarantees that need to be upheld by AI systems will be written in this language, ensuring controllability, legibility, and security.

**Enforcement of AI systems guarantees:**

Lastly, guarantees are only valuable if they are actually enforced. So secure, controllable AI development requires the ability to ensure that the guarantees specified in the previous conditions are actually followed by a given AI system.

This is not the case in current machine learning systems for two reasons. First, as mentioned above, current AI developers are unable to predict how ML systems will behave, even after they have finished training. Thus even after the fact, current machine learning theory provides no way to verify that the AI system follows the specification. And second, current training techniques in machine learning search exclusively for algorithms and AI systems that score high on a set of performance measures. We lack any suitable definitions or specification of control, legibility, or security that can be used as goals of machine learning training processes. This means that ML systems are incentivized to disregard each and any of these properties if that helps them to perform better on their performance indicators or downstream tasks.

Whereas modern ad-hoc safety efforts attempt to fix issues after the fact, playing a losing game of Whack-A-Mole, a responsible approach to building AI must bake in the guarantees in the architecture and the structure of the AI systems themselves. Just like in any other high-risk industry like aviation or energy production, AI research must develop ways to ensure that AI models are industrially “safe by design” against anticipated stresses, as well as building in robustness against unexpected incidents and adversaries’ efforts to attack those models.

## Conclusion

We appreciate this opportunity to provide input on the OSTP RFI. We welcome the opportunity to provide our perspective to OSTP in support of its mission for the American people, and would also like to thank NSF and NITRD for their work to enable public comment. We hope our suggestions are helpful as you develop additional materials, and would be pleased to be a resource and to answer any questions you may have as you move forward.

Sincerely,

David Kasten
Policy and operations
ControlAI