14 March 2025

Submitted by email to:

Faisal D'Souza
NCO
2415 Eisenhower Avenue
Alexandria, VA, 22314
ostp-ai-rfi@nitrd.gov

Response to Office of Science and Technology Policy RFI, "Request for Information: Development of an Artificial Intelligence Action Plan"

ControlAI welcomes this opportunity to reply on the Office of Science and Technology Policy Request for Information (RFI), "Request for Information: Development of an Artificial Intelligence Action 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 AI Action 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 response to inform OSTP's forthcoming AI Action 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 2-4 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.

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. 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

[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

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.**

We must take a different path, one that keeps America in control.

Our recommendations

In November 2023, the US government established the US AI Safety Institute (AISI) to advance our understanding of advanced AI and protect the American people from frontier AI threats. We recommend **passing permanent statutory authorization for a renamed AISI, the US AI *Security* Institute, as a powerful and independent AI security and regulatory agency** outside the Department of Commerce, with the authority to regulate, oversee, and enforce safety standards for frontier AI models.[6] This regulator would ensure that companies developing AI models above certain compute thresholds and general intelligence benchmarks comply with rigorous safety protocols. This would allow the US to harness the benefits of practical AI while mitigating risks posed by the uncontrolled development of superintelligent AI.

The regulator would be led, as is the case for many regulators in the US, by a multi-member panel of regulatory commissioners with staggered terms of office, nominated by the President and confirmed by the Senate. The membership would be roughly equal between the two parties, with no more than half-plus-one of the members from each party. The President would designate one commissioner as Chairman. (This model is similar to, e.g., the US Securities Exchange Commission).

To align incentives and to ensure prompt regulatory action, AISI could be funded through fees from those regulated by it, similar to the FDA's proven approach to industry partnership.

Key aspects of the AI regulator mandate would include:

- Licensing frontier AI developers to ensure AI models are safe before, during, and after development.
- Prohibiting the development and/or deployment of dangerous AI capabilities, such as unauthorised replication, environmental breakout, and autonomous self-improvement.
- Oversight of high-computation AI models and applications that present catastrophic or extinction level risks.

[6] Note: there are a wide range of costs and benefits to moving AISI out of the Department of Commerce. Policymakers might feasibly choose instead to retain it in its current location, locate it within another government Department (e.g., Energy, Homeland Security), or create a CFIUS-like coordinating body.

There are meaningful legal considerations (e.g., US Persons rules) that mean that AISI would have difficulty interfacing with US AI companies if it was located in the Department of Defense or the Intelligence Community, so we specifically recommend against those options.

- Establishing security and product safety standards for the design, development, deployment, and monitoring of AI systems.

- Threat, scenario, and trend assessment of security and risk impacts of AI.

The Licensing Framework

At the core of the regulator's power would be a **three-tiered licensing system** aimed at managing the development and deployment of frontier AI models above critical compute thresholds. These licenses would ensure that only AI developers and operators that meet safety requirements can proceed with their work.

1. **Training License**
   For AI developers aiming to train models that exceed a set computational power threshold, set at 10^25 FLOP[7]. Inspired by other current industrial security and safety boards, regulated applicants must present detailed risk mitigation plans for managing developing and deploying AI for their intended use, including shutdown procedures for AI systems that pose unacceptable levels of risk.

2. **Compute License**
   Required for cloud service providers and data centres operating above 10^17 FLOP/s. The compute license would ensure that large-scale computational power is not misused for unregulated AI development. Licensees must implement hardware tracking and know-your-customer (KYC) requirements to maintain transparency and security over computing resources.

3. **Application License[8]**
   For developers seeking to develop applications using a licensed model; they would have to declare the purpose or purposes and sector or sectors for which the model would be used. This system would also ensure that modifications to approved AI models remain compliant with safety regulations, particularly when model capabilities are enhanced. Automatic approval would apply to new licensing submissions where they were substantially similar to existing licenses with no significant capability upgrades, though models would still need to seek relevant sector-specific approvals.

Prohibiting Dangerous AI Capabilities

The regulator would have the power to enforce prohibitions on **specific high-risk AI behaviors**, ensuring that even models operating below regulatory thresholds do not engage in hazardous activities. (These prohibitions are, in part, inspired by the unacceptable risk

[7] Floating-point operations
[8] Some Congressional leaders have proposed that AI should be regulated via a sector-specific approach. The application license system would enable such an approach by ensuring that unscrupulous AI companies could not evade sector-specific regulation by developing a dangerous general-purpose model, but only seeking regulatory approval for one narrow sector-specific use.

thresholds that governments have committed to identify in the Seoul agreement). These prohibited capabilities should include:

- No Superintelligent AIs: AI must not surpass human intelligence in general tasks.
- No Unbounded AIs: AI systems should not be developed or deployed *unless* a robust safety case can be made regarding their capabilities of concern, ensuring AIs remain predictable and controllable.
- No Environmental Breakout: AI systems must not escape their designated environments or access external systems or networks, even with authorisation, if the regulator deems the degree or scope unsafe by design. (E.G., this ensures that AI models do not pose cybersecurity threats, nor become capable of evading security measures placed upon them.)
- No AIs Improving AIs: AI systems should not improve or develop other AI systems, particularly those not directly written by humans, to prevent runaway AI development that humans cannot control.

Oversight of High-Computation AI models

Through the licensing process, AISI would be able to proactively and iteratively monitor AI model development and deployment, and ensure that the President was fully apprised of all AI capabilities and their implications for US national security.

Establishing Security and Product Safety Standards

To enable lighter-weight regulation, AISI would have the legal authority to establish common standards for AI models' development and deployment while protecting Americans' security and product safety, in line with any other high-potential industry's standards. Such standards would include, but are not limited to, cybersecurity and physical security, insider threat, and reliability standards.

Threat, Scenario, and Trend Assessment

AISI would build in-house capabilities to understand the security and other risks of AI development and deployment, which it could also use to support US Intelligence Community and DoD analysis.

Governance and Flexibility

The regulator's governance would remain flexible to adapt to future AI developments and risks with the establishment of an **AI Safety Board**, which would be responsible for defining key regulatory thresholds and capabilities for licensing requirements, and have the power to **order the shutdown of dangerous AI models or applications.** Such a board would be staffed by experts and also solicit industry, IC, and DoD perspectives.

A newly created **Scientific Advisory Group** would provide expert input on emerging AI capabilities, risks, and safety measures. This advisory group, drawing from the best experts at

world-leading American institutions like MIT, Harvard, Yale, Stanford, and others, would work closely with the Board to ensure that regulatory decisions are scientifically informed and aligned with global safety standards.

Looking towards the future – international deals

Beyond these initial steps to protect America from the risk of uncontrolled, powerful AI systems, we recommend building on national-level action by leading internationally as well. The Trump 45 Administration found ways to bring America's friends, rivals, and adversaries to the negotiating table and make transformational deals. It is no secret that the AI policy world has been talking widely about hopes for the dealmakers of the Trump 47 Administration to craft what would be, candidly, the biggest deal in human history – an international deal to manage the risk of AI by getting other countries (e.g., the UK, France, China) to accept American standards for risk-reduction, impose equivalent regulations on their own companies and governments, and pay their fair share of the cost of AI security by funding research and security efforts. Such a deal would ensure the safety of every American and everyone else on the planet; it would be bigger and more important than any effort to date to curb the risks of war or nuclear proliferation that has won the Nobel Peace Prize. We hope the Trump Administration aggressively pursues such an effort.

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
dave@controlai.com