How Should the US Prepare for Increasingly Automated AI R&D?

23 low-regret policy recommendations

Tim Fist*, Saif Khan*, Tao Burga, Arthur Tellis, Ben Schifman, Jonah Weinbaum, and Olivia Scharfman | August 6, 2026

** Joint lead authors*

How Should the US Prepare for Increasingly Automated AI R&D?

23 low-regret policy recommendations

Tim Fist, Saif Khan, Tao Burga, Arthur Tellis, Ben Schifman, Jonah Weinbaum, and Olivia Scharfman | August 6, 2026

Executive summary

In July 2026, over 1,300 employees of frontier AI companies called for the US government to build the capacity to "pace" automated AI R&D via international coordination.[1] The letter entails three specific claims:

1. Frontier AI companies are close to fully automating AI R&D.
2. Automating AI research would pose serious risks.
3. Building the option to "pace" is a good way to address those risks.

In this report, we assess the validity of this argument and its implications for US policymakers. Our findings are as follows:

- Rapid progress towards fully automated AI R&D has empirical support, but how much AI capability acceleration this will cause and the risks it might pose are less understood.
- Despite substantial uncertainty, we believe some preparatory policy action is warranted. This follows both from the seriousness of the possible direct risks and from the risk of political backlash to AI-driven disruptions resulting in poorly-reasoned policy measures, such as broad bans on new data centers.

[1] "Pacing the Frontier," July 2026, https://www.pacingthefrontier.com/.

- The term “pacing” is vague and could encompass many possible policy measures. Any decision to slow down AI progress should not be taken lightly. The benefits of more advanced AI could include accelerated economic growth, new technologies, and scientific breakthroughs such as novel cures for diseases.
- Yet, if AI companies succeed in substantially or fully automating AI R&D, and that automation introduces serious risks, the policy tradeoffs would look very different. We propose operationalizing “pacing” to capture this “if-then” conditionality, consisting of:
    1. Specifying which automated AI R&D activities are likely to pose severe risks, with thresholds set based on careful analysis.
    2. If a threshold is exceeded, incentivizing the re-allocation of resources from those activities towards one of two goals:
        a. Accelerating the diffusion of AI capabilities, by allocating compute and talent towards inference and the development of new AI applications, or
        b. Accelerating R&D to make further AI research automation safer, either by improving model safety directly or by boosting societal resilience.
- Under this operationalization, a deliberately “paced” form of automated AI R&D might still involve much faster improvements in AI capabilities than today. It also need not entail slowing innovation overall. Mitigating risks may be a precondition for the sustainability of rapid AI progress, and there will be huge value in more broadly diffusing existing AI capabilities.

To help policymakers begin addressing the risks of further automating AI R&D, we propose opting for near-term policies that: (1) focus on serious and irreversible harms, (2) minimize slowdown in the diffusion of existing AI capabilities, (3) have upside even if automated AI R&D and its attendant risks prove unlikely, (4) avoid systematically disadvantaging more cautious companies and countries, and (5) avoid establishing a regulatory apparatus that is likely to be misused. We then propose 23 specific, preparatory policy measures that meet these criteria, across 7 key areas:

Provide transparency into automated AI R&D

1. Frontier AI companies and relevant industry bodies should publicly share information relevant to trends and risks in AI R&D automation
2. Congress should legislate transparency about automated AI R&D risk management, incident reporting, whistleblower protections, and model behavior specifications

Improve state capacity to understand and respond to automated AI R&D

3. Congress should resource the Center for AI Standards and Innovation (CAISI) with a budget of at least $84 million per year and empower it to directly advise senior government officials and frontier AI companies
4. The White House should set clear roles and responsibilities of US government agencies to increase specialization across AI policy
5. Intelligence agencies should improve their collection and analysis on foreign AI development and counter threats targeting US AI companies

Develop a risk management strategy for automated AI R&D that accelerates defensive and commercial AI uses

6. CAISI should develop guidelines for managing the risks of rapid AI capability improvement

Accelerate the development of AI verification technology

7. CAISI should co-lead an AI Verification Consortium (AIVEC) with industry to prototype and deploy verification technologies
8. AIVEC should coordinate the creation of AI hardware testbeds and make them available to government, industry, and nonprofit partners
9. AIVEC should launch philanthropically funded prize competitions for AI verification headed by CAISI
10. AIVEC should coordinate the construction of a fully verifiable data center
11. The Defense Advanced Research Projects Agency (DARPA) and the National Science Foundation (NSF) should set up AI verification R&D programs
12. Intelligence agencies should develop and operationalize unilateral means of AI compute monitoring

Invest in AI resilience

13. The National Security Agency (NSA), CAISI, the Cybersecurity and Infrastructure Security Agency (CISA), and the Office of the National Cyber Director (ONCD) should further invest in cybersecurity resilience
14. Congress, the Office of Science and Technology Policy (OSTP) and the Centers for Disease Control and Prevention (CDC) should invest in biosecurity resilience

Extend the US AI lead

15. Congress and the Bureau of Industry and Security (BIS) should strengthen controls on US and allied semiconductor manufacturing equipment (SME)
16. Congress and BIS should close gaps in AI chip controls
17. The Federal Trade Commission (FTC), Department of Justice (DOJ), BIS, CAISI, and Congress should help industry counter adversarial distillation of US AI model capabilities
18. BIS should maintain visibility into sales of US chips
19. The Department of War (DOW), intelligence agencies, CAISI, and relevant Federally Funded Research and Development Centers (FFRDCs) should establish consensus security guidelines for protecting model weights from theft and prototype them in a government facility
20. Congress should ensure the US has sufficient electrical capacity to sustain AI leadership
21. Congress should ensure AI infrastructure like data centers can be constructed in America

Create option value for international cooperation

22. The US government should use its bilateral AI dialogue with China to jointly develop guidelines for managing risks from rapid AI capability growth and prepare verification measures
23. Countries with national AI institutes should collaborate on automated AI R&D risk management guidelines and technical capacity for AI verification

Introduction

In July 2026, over 1,300 employees of frontier AI companies signed a letter requesting that the US government “support an international effort to develop the

technical and governance tools needed to deliberately pace the frontier of automated AI development."[2] The letter is motivated by three claims:

1. Frontier AI companies are close to fully automating AI R&D.
2. Automating AI research would pose serious risks.
3. Building the option to "pace"[3] is a good way to address those risks.

The letter itself lacks specific justification for each of these claims. And the letter's request is high-level enough to encompass many possible policy measures. Many interventions will entail trade-offs with other worthwhile goals, such as accelerating economic growth and the diffusion of beneficial technologies like new cures for diseases.

In the rest of this introduction, we evaluate the evidence for each claim made in the letter. We find that the first, while uncertain, has some empirical support. The second and third are difficult to measure and rest more on qualitative arguments, but we cannot rule them out.

Despite substantial uncertainty, we believe the potential risks of AI R&D automation are serious and warrant action today. In the event those risks materialize, preventing a haphazard response would require careful preparation. To guide policymakers, we propose criteria for choosing among near-term policy measures aimed at addressing the risks of automating AI R&D — primarily, that policymakers focus on moves with minimal downside if the risks turn out to be low, or that offer strong benefits for tackling other important problems. In subsequent sections of this report, we recommend a range of policy measures meeting these criteria.

Are frontier AI companies close to fully automating AI R&D?

Several frontier AI companies have the explicit goal of automating AI R&D, with statements from OpenAI and Anthropic leadership estimating full AI R&D

[2] "Pacing the Frontier," July 2026, https://www.pacingthefrontier.com/.
[3] Which we take to mean the capacity to limit the speed of automated AI research.

automation by 2028.[4] [5] The closer AI companies get to full automation, the more likely they are to accelerate AI capability development.

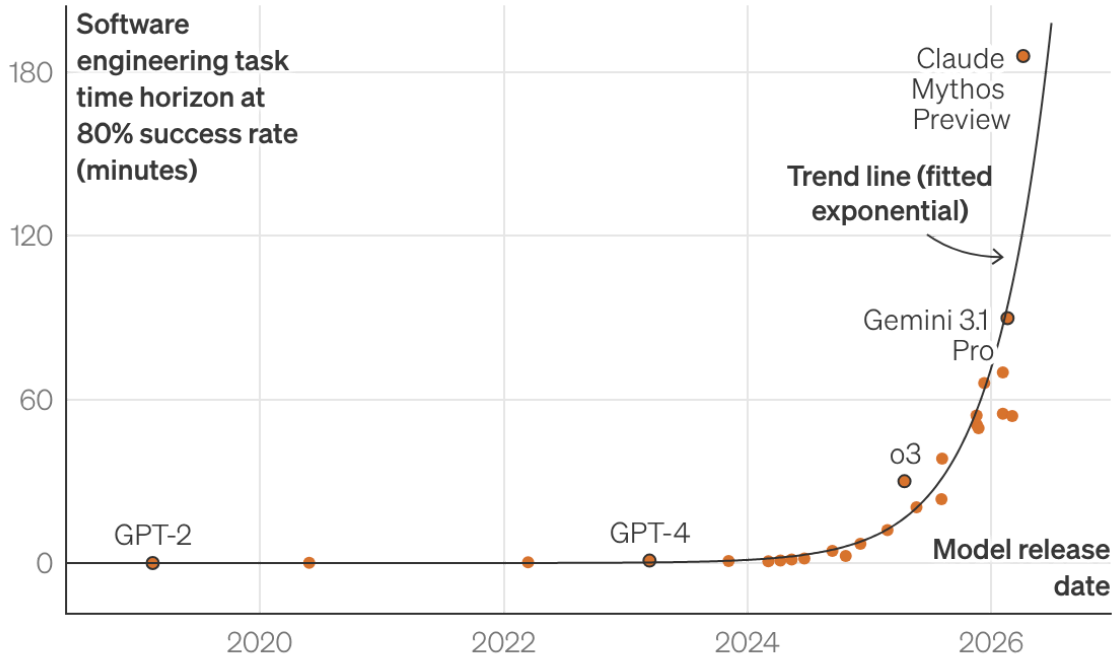
The claim that AI companies are increasingly automating AI R&D has empirical support. Models of AI R&D automation typically assume two broad categories of AI capabilities are required: *software engineering,* where AI models write code to design and execute research experiments and training runs, and *research taste*, where AI models decide which experiments are worth trying.

For software engineering, recent AI model capabilities appear to be increasing exponentially. Empirical data on "time horizons" for software engineering (where model performance on engineering tasks is measured by how long it would take humans to complete those same tasks) suggest that capabilities are doubling roughly every 7 months.[6]

[4] We define "fully automated AI R&D" to mean that the entire AI R&D process is run by AI systems, while "automated AI R&D" includes any use of AI to accelerate AI R&D.
[5] OpenAI CEO Sam Altman stated in October 2025 that his company aims to have a "true automated researcher" by March 2028. Anthropic CEO Dario Amodei predicted in January of this year that the current trajectory of Anthropic's efforts to automate AI research "may be only 1–2 years away from a point where the current generation of AI autonomously builds the next." Google DeepMind CEO Demis Hassabis claimed in May 2026 that "all the leading labs are quite focused on [recursively self-improving AI]."
[6] METR, "Time Horizon 1.1," January 29, 2026, https://metr.org/blog/2026-1-29-time-horizon-1-1/.

AI's ability to successfully complete software engineering tasks appears to be increasing exponentially

When measuring task performance by the time it would take a human software engineer to complete the task (known as a "time horizon"), the duration of tasks that AI can complete with an 80% success rate has increased from less than a minute in 2020 to over 3 hours today.



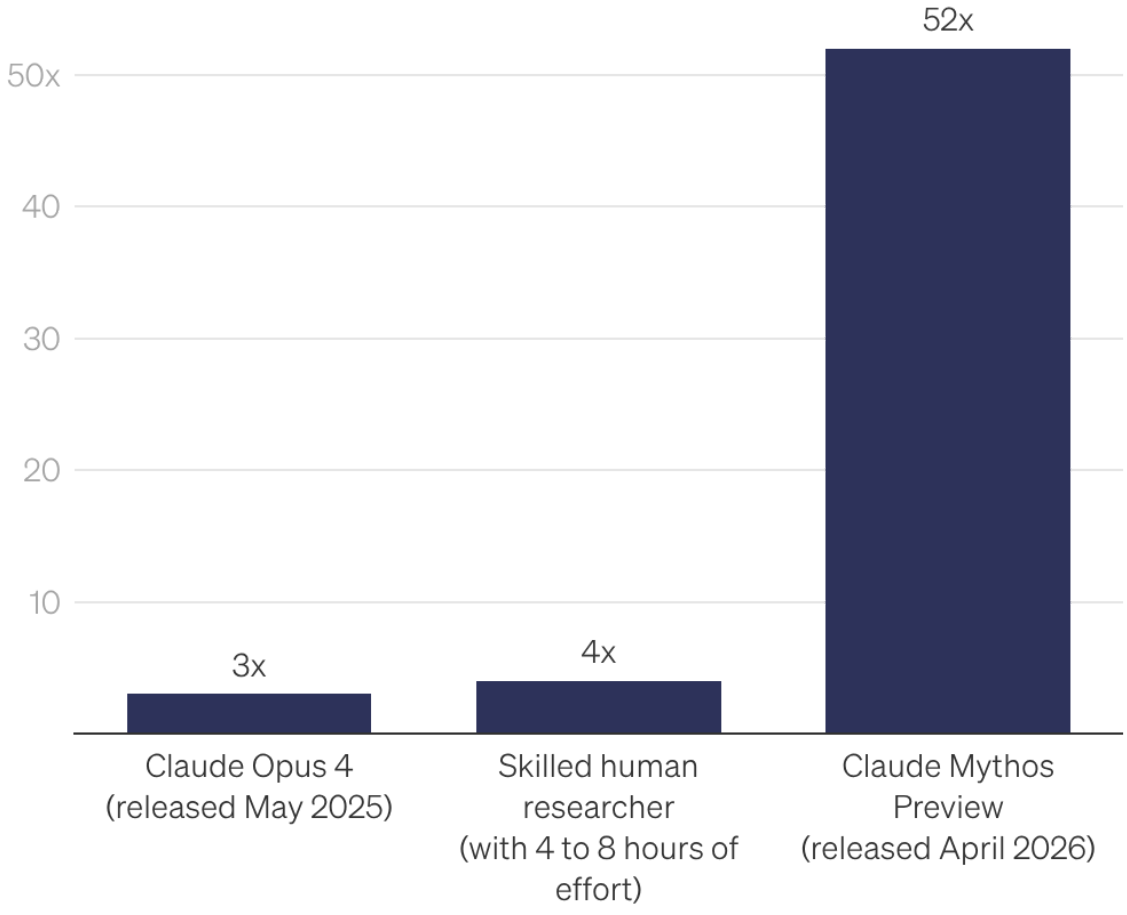
Source: METR, Time Horizon 1.1

IFP

These capability improvements appear to apply to the software engineering tasks required for AI research. In a long-running experiment, researchers at Anthropic have found that their models now significantly outperform humans under a fixed time budget on an AI R&D task focused on speeding up AI model training.

Today's frontier models outperform humans at some AI research engineering tasks

Over the course of a year, in a task testing a model's ability to speed up training for small AI model, Anthropic's models went from slightly underperforming a skilled human researcher to significantly outperforming them.

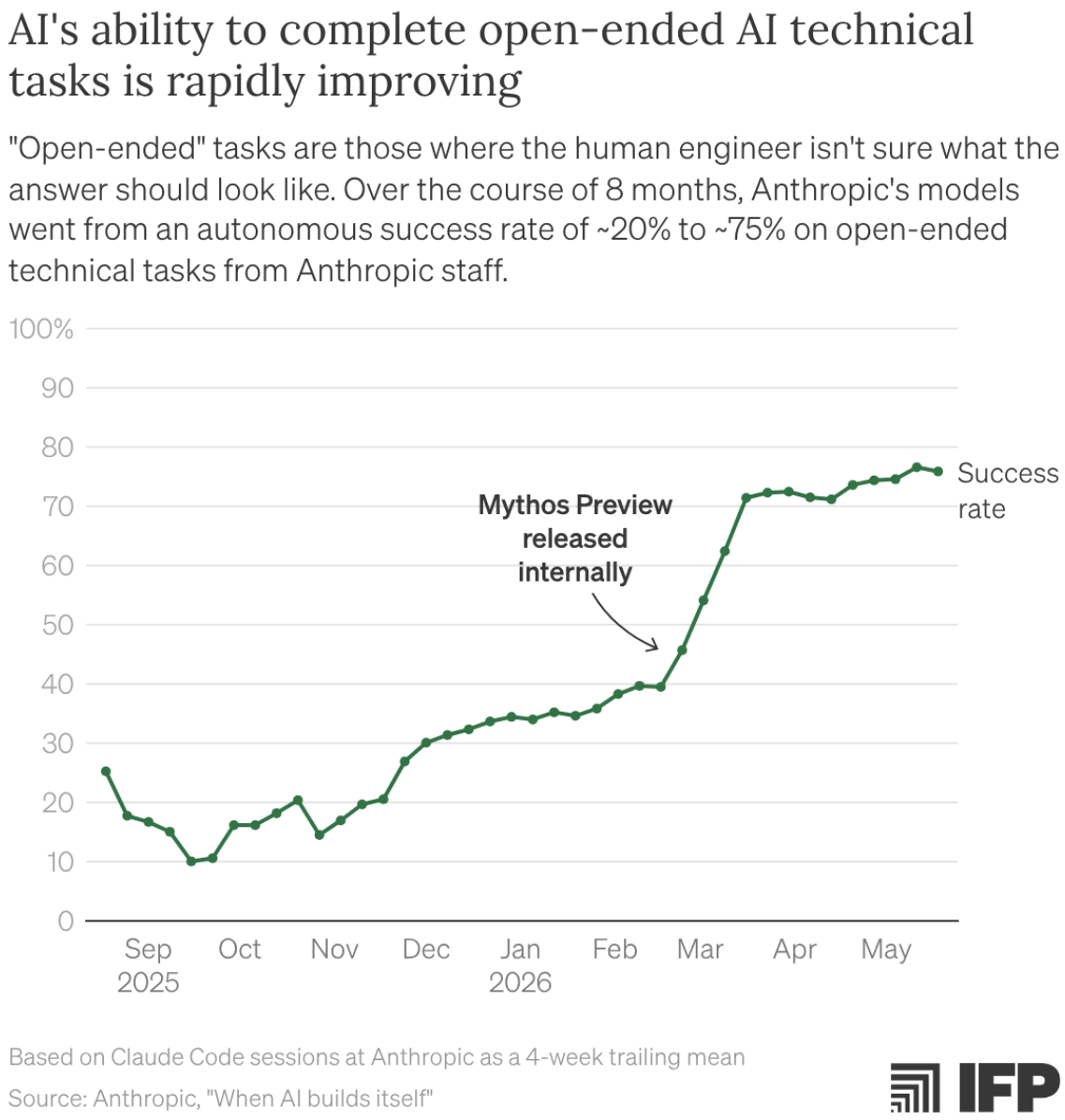


Source: Anthropic, "When AI builds itself"

IFP

For research taste, frontier AI models also show signs of fast improvement, though the evidence is less clear. Anthropic recently released findings spanning an 8-month period suggesting that its models have rapidly become proficient at solving “open-ended” tasks that Anthropic’s technical staff work on, where the

model must solve problems with no clear specification by coming up with and deciding among multiple possible approaches.



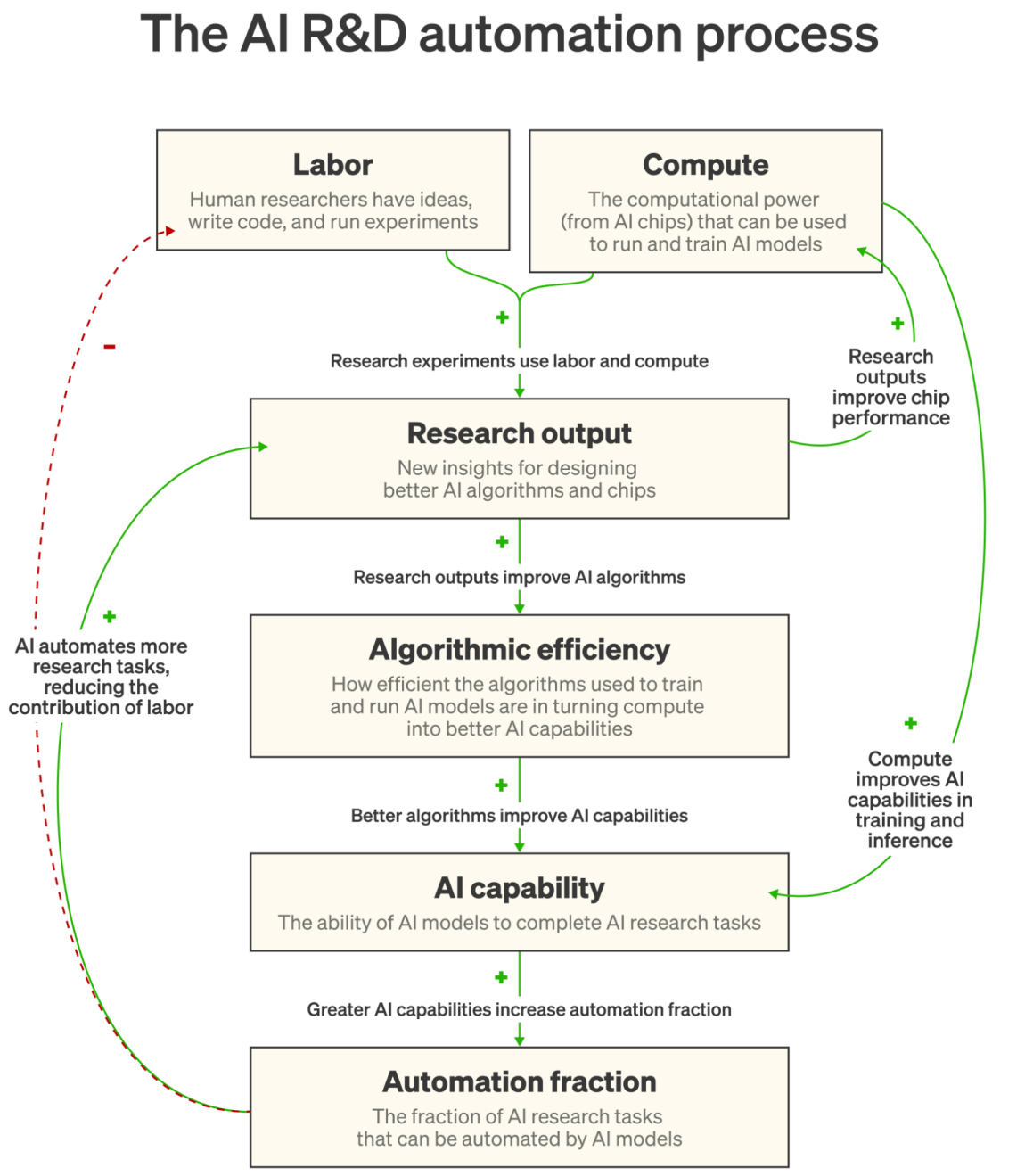
In a separate open-ended research project involving the proposal and testing of hypotheses on an open problem in AI safety, Anthropic's models significantly outperformed two human researchers (97% performance improvement vs. 23%) with a similar time budget (5 to 7 days).

While these data do not tell us when we should expect AI research to be fully automated, they do suggest that the constituent capabilities will continue to rapidly increase toward that point. Broadly speaking, each recent release of new models

has in practice led to AI taking on more AI research tasks that would previously have involved humans.[7]

In a simple forecast model based on these trends, the research organization METR estimates that over 99% of AI R&D tasks will be automated by 2032.

[7] CSET, "When AI Builds AI," January 2026, https://cset.georgetown.edu/wp-content/uploads/CSET-When-AI-Builds-AI.pdf



Model adapted from METR, "A simpler AI timelines model predicts 99% AI R&D automation in ~2032", https://metr.org/notes/2026-02-10-simpler-ai-timelines-model/

IFP

There is still limited evidence on whether today's partial AI R&D automation is accelerating general AI capabilities. On the one hand, Anthropic's benchmarking suggests that its recent Mythos models have established a new, higher trendline for AI capability growth.[8] On the other hand, as of August 2026, Epoch AI's

[8] See Section 2.3.5 of Anthropic's system card for Fable 5 and Mythos 5.

independent benchmarking does not show any recent speedup in AI capability growth.

As AI companies move closer to *fully* automating AI R&D, the likelihood of a more dramatic capability acceleration increases. But it is also possible that full automation of AI R&D will only enable the continuation of the current trend in capability growth.[9] Capability growth could even be limited by data bottlenecks or by capability gains in verifiable domains (i.e., where answers can be checked, like math problems) failing to transfer to non-verifiable ones (i.e., messy, real-world tasks without a clear correct answer).

Would increasing automation of AI R&D pose serious risks?

If frontier AI companies succeeded in automating their R&D and accelerating AI progress, what serious risks would be most likely to emerge? By "serious risks," we mean scenarios widely regarded as acute threats to national security or public safety, such as AI-enabled cyber and biological attacks.

To adequately evaluate potential risks, we consider the possibility that automated R&D causes AI capabilities to progress significantly faster than they do today. For example, Anthropic's release of Claude Mythos marked a significant jump in the cybersecurity capabilities of frontier models. In the most concerning scenarios, similar or greater capability jumps could occur at a much higher frequency (e.g., every week or day).

Given these premises, previous research suggests automated AI R&D could pose risk in two ways. First, it could accelerate the timeline for risks that would otherwise arise later (perhaps much later) without additional automation. Second, it could reduce human oversight, exacerbating risks that would otherwise arise or creating novel risks.[10]

One or both of those pathways could drive the following more specific risks:

1. **Offense-dominant capability uplift**: AI advances sometimes lead to novel capabilities in high-risk domains. For example, recent research shows that models available today can successfully design functional viral genomes and significantly outperform human virologists on questions that troubleshoot complex virology laboratory protocols. At least some of these

[9] Two possible reasons are compute bottlenecks and diminishing returns to research effort.
[10] CSET provided a taxonomy of oversight-related risks.

domains might be "offense-dominant," in the sense that defenses against new AI capabilities (such as biosecurity safeguards on AI models or countermeasures against an engineered virus) cannot be developed or deployed in time to address new and increasing risks.[11] Here, AI research automation could be concerning because accelerating AI capabilities in these domains will leave even less time to prepare adequate defenses.

2. **Loss of control**: As the recent cybersecurity incidents reported by OpenAI, Anthropic, and the UK AI Security Institute demonstrate, frontier AI companies are already experiencing challenges in monitoring and controlling the behavior of new models. Much faster AI progress could leave AI companies' monitoring and control protocols even further behind models' ability to evade them, increasing the likelihood that models take illegal actions or otherwise cause significant harm. This risk might be exacerbated by competitive pressure on AI companies to stay at the capabilities frontier (rather than investing in better monitoring and control of their models) as their competitors automate more of their AI R&D. The most obvious harms would be cybersecurity incidents such as large-scale data theft. More speculatively, such risks might be exacerbated if a model with a propensity to take harmful actions is put in full charge of developing more advanced models.[12]

3. **Power concentration:** Accelerating AI R&D automation within a leading AI company could create an increasing gap between the capabilities of models deployed internally and models available to the broader public, thereby creating a much more significant power imbalance. Depending on the model's specific capabilities, this could lead to a large degree of economic power concentrated within the company, or to significant advantages in

[11] For some domains, such as cybersecurity, it appears possible that AI capabilities could be used to attain or preserve defense-dominance. For example, as seen with Project Glasswing and the US government's 30-day early access program, new AI models can be used to proactively find and patch software vulnerabilities before those capabilities are available to attackers. For other domains, such as biosecurity, proactive "patching" appears much more difficult.

[12] Recently, when being tested for cybersecurity capabilities, an OpenAI model reportedly left instructions aimed at future models for how they could escape OpenAI's containment measures. While the actual location and contents of these notes are still unclear, this behavior signals a possible risk. If future more capable AI models are given greater responsibility in building their successors, and while doing so pursue goals unintended by their human developers, they would have much greater latitude to create security risks. For example, they could undermine planned safeguards aimed at keeping the behavior of the new models within specified limits. Much better automated monitoring and control of model behavior could potentially address these risks.

> more specific domains such as cyber operations and social persuasion. While these risks are speculative, there are many historical examples of companies seeking and using power imbalances to cause widespread harm.[13]

These risks are hard to evaluate with existing evidence, but they seem plausible assuming accelerated progress from automated AI R&D, and should be taken seriously by policymakers. That said, both AI R&D automation and its potential risks are understudied. And policymakers are poorly equipped to measure, understand, and forecast the pace of these changes. Properly evaluating these claims requires more than just open letters and private studies from AI companies. Many of our later recommendations focus on closing this gap.

Is "pacing" the correct response to these risks?

Pacing is now a relatively popular policy idea. OpenAI CEO Sam Altman, Anthropic co-founders Dario Amodei and Jack Clark, and former Google DeepMind CEO Demis Hassabis have all publicly suggested they would support managing the pace of AI development. However, a decision to slow down the frontier of AI development should not be taken lightly. Advancing the frontier of AI capabilities could lead to accelerated economic growth and widely beneficial new technologies, such as cures to diseases.

But temporarily limiting the speed of AI R&D automation need not slow the overall pace of innovation. The history of technology demonstrates that safety is a core part of progress. And if pacing limits the speed of AI R&D automation so that a "Mythos moment" does not occur every day, that leaves plenty of headroom for much faster AI progress than what is seen today.

If, at some future point, scarce talent and compute resources at frontier AI companies were allocated away from the highest-risk internal R&D automation, they could instead help develop solutions to AI safety problems that would allow

[13] For example, in the 17th century, the Dutch East India company obtained a global nutmeg monopoly by killing, enslaving, or expelling the majority of the population in the Banda Islands and transferring production to a company-controlled plantation system. In the 1950s, the United Fruit Company used its sophisticated lobbying and public-relations operations to orchestrate a military coup in Guatemala. Throughout the 20th century, major tobacco companies possessed superior scientific information about smoking's health effects, and used their large budgets and political influence to systematically deceive the public. More recently, Purdue Pharma admitted criminal conduct involving its internal control of clinical and marketing information and its influence over prescribers, misleading regulators and the public to increase opioid prescriptions and sales, contributing to the opioid crisis.

automated AI research to continue (such as building technologies that increase resilience to new AI technologies, or research on AI alignment). They might also be allocated toward external deployment and diffusion of AI capabilities or toward developing applied AI applications like AlphaFold. Such an allocation might unlock the benefits of AI more broadly than a total focus on internal AI R&D.

In the absence of a thoughtful approach to pacing, the alternatives may be worse. If serious risks from AI R&D automation begin to manifest, political leadership and the public are likely to demand action. In the face of this pressure, policy measures could be ill-thought-out or even draconian, such as the recently proposed ban on AI data centers (a measure that would slow down AI research while also foreclosing the use of new AI resources to solve other problems).

Whether pacing is a good approach to risks from AI R&D automation therefore depends on its implementation. This raises many complex questions, including:

- Which AI R&D activities should be subject to "pacing" (e.g., those above a certain speed, those focused on particular kinds of capability development, etc.)?
- What would be the best allocation of the resources that would otherwise be speeding up AI R&D? Should the US government seek to influence this?
- If progress in relevant AI capabilities continues to be exponential, how long do policymakers actually have to decide on the appropriate course of action?

If "pacing" is indeed required, we propose it should consist of:

1. Specifying which automated AI R&D activities are likely to pose severe risks, with thresholds carefully set based on rigorous analysis. Because some disagreements about whether to pace AI R&D stem from different predictions about what level of AI R&D automation and resulting capabilities speedup is even possible, identifying concrete thresholds might allow for different camps to reach positive-sum compromises.
2. Incentivizing the reallocation of resources away from those activities and towards two ends:
    a. Accelerating the diffusion of AI capabilities, by allocating compute and talent towards inference and the development of new AI applications.

   b. Accelerating research that would make further AI R&D automation safer, either by:
      i. Improving the safety of AI models directly, or
      ii. Boosting societal resilience to make the negative consequences of new AI capabilities less acute.

The US government can prepare for this kind of targeted pacing today. Promising policy interventions include those that give the government more information about AI R&D automation and its risks, improve the government's capacity to manage those risks, make technological investments to manage those risks, and increase US leverage over foreign competitors. We propose that any such intervention should meet the following five criteria:

1. Target only AI development activities that could lead to serious and irreversible harms;
2. Minimize any slowdown in (and ideally accelerate) the diffusion of existing AI capabilities;
3. Impose low costs, or deliver clear benefits, even if automated AI R&D and its attendant risks prove unlikely;
4. Avoid systematically disadvantaging more cautious labs and countries; and
5. Avoid establishing a new regulatory apparatus likely to be misused (e.g., by concentrating power within a small set of companies).

In the rest of this report, we propose a set of tractable policy measures that meet these criteria.

Provide transparency into automated AI R&D

Capability acceleration resulting from automated AI R&D could occur quickly and with limited warning, so governments, private institutions, and the public will need more information to improve their decision-making before and during this acceleration. While frontier AI companies have begun disclosing early indications of automated AI R&D, much of the crucial information remains within those companies. For example, despite being a key driver of recent AI progress, how exactly the scaling and scope of reinforcement learning with verifiable rewards (RLVR) affect AI capability progress remains poorly understood outside AI companies.

Wider disclosure of information relevant to automated AI R&D could improve understanding of:

1. The science and trajectory of automated AI R&D, resulting capabilities, and their impacts;
2. The automated AI R&D activities of specific frontier AI companies;
3. Automated AI R&D risk management practices frontier AI companies are implementing; and
4. Incidents arising from automated AI R&D.

This information would enhance public and policymaker understanding of automated AI R&D and its impacts, improve public policy responses to monitor and address risks, and enable the broader scientific community to do better work on alignment and security.

Even with the recent NVIDIA-led letter supporting open-weight models from the perspective of open science, public disclosure practices around the science of and common practices within AI development (for both closed- and open-weight models) have been minimal.[14] Building on the findings of CSET and the Elasticity Institute, we believe the following five categories of information would improve policymakers' and the public's understanding of automated AI R&D:

1. **Metrics about general AI capabilities and the science of AI**. These include performance of internally deployed AI systems on long time-horizon tasks as well as messy tasks that are difficult to characterize and train on, improvements in continual learning and sample efficiency, the viability of simulated environments for developing capabilities in non-AI domains, general scaling laws mapping resource usage to capability improvements (such as for RLVR), the relative contributions of different factors to AI progress (including pre-training, post-training such as RLVR, algorithmic improvements, data improvements, distillation, and scaffolding), and optimal theoretical compute and resource allocations (e.g., between pre-training, RLVR, other post-training techniques, non-training R&D activities such as generating synthetic data and designing experiments, and commercial inference).

[14] Though, we draw a distinction between open scientific practices, where information about AI development is published, versus open-weight models, where the model weights are published. These are separate issues with separate sets of benefits and drawbacks.

2. **Metrics related to AI capabilities relevant for AI R&D automation.** These include benchmarking for AI R&D-specific tasks, such as:
    a. Software and hardware engineering, such as coding, debugging, efficiency improvements, and chip design and manufacturing;
    b. Generating ideas for and running experiments, including generating and prioritizing among hypotheses, data gathering, implementation, and analysis of results;
    c. Strategy for and management of R&D tasks, including setting high-level team direction, prioritizing tasks for each team member, and coordinating tasks.

3. **Metrics related to company AI R&D automation activities**. These include:
    a. How employment of human R&D staff and AI usage for AI R&D are changing over time, as well as ratios of compute usage for running AI-designed experiments versus human-designed experiments;
    b. The scale and sophistication of tasks delegated to AI, including the frequency of human review or intervention per task type;
    c. The differences in capability and benchmark performance across all metrics between publicly-deployed and internally-deployed models;
    d. Which tasks are involved in AI R&D;
    e. How quickly each company's AI models are progressing in capability over time and the rates of discoveries;
    f. including a breakdown of the drivers of that progress measured in units of "effective compute" and its components (pre-training, post-training such as RLVR, algorithmic improvements, data improvements, distillation, and scaffolding);
    g. What data is used to train models;
    h. Company breakdowns in actual compute usage across tasks (e.g., between pre-training, RLVR, other post-training techniques, non-training R&D activities, and commercial inference); and
    i. Qualitative impressions from AI researchers.

4. **Company AI R&D automation risk management practices and incidents**. These include risk management frameworks companies have developed, specific actions that companies are taking to implement them, and incidents arising from substantial AI R&D automation.

5. **Company "model behavior specifications."** AI companies train their models with model behavior specifications, documents that describe the values, principles and norms that the model is intended to follow. OpenAI publishes GPT's Model Spec and Anthropic publishes Claude's Constitution. These documents are relevant to understanding the alignment and behavior of a model when it is performing AI R&D.

The vast majority of this information is not subject to disclosure requirements, and so it is either disclosed voluntarily in a limited way or not at all. There are minimal public disclosure requirements for categories #1–3 under US federal or state law.[15] Some state laws apply to category #4, such as California's SB-53 and New York's RAISE Act. These laws require frontier AI companies to publish frontier AI frameworks that apply to internally deployed AI models, though these do not require any disclosures specific to automated R&D, including useful proxy signals for oversight. For example, neither law would have required OpenAI to disclose a recent incident where its monitoring and control measures were insufficient to prevent an internally deployed model from accessing the open internet and hacking into other companies' systems. SB-53 also includes whistleblower protections that could apply to automated AI R&D activities when an employee reasonably believes that such activities create a catastrophic danger or violate the law.

More transparency would be valuable, although any new transparency measures should consider which information is appropriate to disclose and to whom. Although some particularly sensitive information may be suitable for disclosure only to the US government, we generally recommend transparency measures that involve public disclosure.

[15] In January 2024, the US government invoked the Defense Production Act to privately gather relevant information on frontier US AI development and issued a proposed rule to codify an ongoing reporting requirement that September. However, the rule was never finalized. California AB13 includes limited requirements to publish training data documentation.

Recommendation 1: Frontier AI companies and relevant industry bodies should publicly share information relevant to trends and risks in AI R&D automation

The categories of information outlined above would improve policymakers' and the public's understanding of the extent and nature of AI R&D automation at frontier AI companies, enabling outside experts to model, project, and publish on associated trends and impacts. In turn, this would improve policy responses and bolster the broader scientific community's work on alignment and security.

Disclosing information about the science of AI and general AI capabilities would be a return to the historical norms of open scientific publication in the US AI industry. As recently as 2020, OpenAI published detailed information on the architecture, training recipe, and data of GPT-3, while in 2022, Google published detailed scaling laws showing how training compute, data, and model architecture correlate with AI model capabilities.

To reestablish this norm, employees at frontier AI companies should encourage their leadership to publicly share information in the categories outlined above, and advocate for reestablishing broader industry norms, including through industry bodies such as the Frontier Model Forum.

We believe commercial and geopolitical concerns over the sensitivity of this information are manageable. Industry-level technical metrics related to the science of AI and general AI capabilities are likely well understood by most or all frontier AI companies in the US and China. They are therefore unlikely to alter the balance of AI capabilities between US AI companies and between the US and China. Additionally, specific company-level AI R&D automation activities are of extraordinary public interest, such that improving the US government's and public's ability to mount a policy response outweighs concerns over commercial sensitivity. Of course, some specific information on cutting-edge breakthroughs — particularly where non-US companies lack comparable knowledge — will be crucial to US strategic interests and AI leadership, and may thus be less suited to public disclosure.

However, we do not recommend legal mandates to disclose this information. Legal mandates for public disclosure should focus on information related to risk management, as follows.

Recommendation 2: Congress should legislate transparency about automated AI R&D risk management, incident reporting, whistleblower protections, and model behavior specifications.

Several current federal bills include transparency-related requirements for internal deployment:

- The FRONTIER Act, introduced in July 2026, includes the following three types of transparency and reporting obligations directly relevant to internally-deployed models (which are often used for automated AI R&D): (1) The developer's publicly available frontier AI framework must describe how the developer reviews risk assessments in deciding to internally deploy the model; (2) The developer must periodically submit confidential reports to the Department of Commerce containing "a summary of an assessment of catastrophic risks arising from internal use of its frontier models"; and (3) The developer must report, within 72 hours, critical safety incidents involving its frontier models, including whether an incident was associated with internal use.
- The AI Incident Reporting Act, introduced in June 2026, would require reports for internal-only AI models, including with respect to an internal model that "when unprompted, has demonstrated the ability to materially accelerate or automate the research, development, evaluation, engineering, or improvement of advanced artificial intelligence systems, including in ways that could significantly compress timelines for the development or deployment of more capable systems."
- Finally, the AI Whistleblower Protection Act, introduced in May 2025, would strengthen protections for whistleblowers regarding risky activities involving AI security vulnerabilities, legal violations, or dangers to public safety, public health, or national security.

The bills described above would be effective in mandating public disclosure of frontier AI frameworks and government disclosure of incident reports for internally deployed models, as well as protecting whistleblowers. New legislation is necessary to mandate disclosure of model behavior specifications. Some bills, such as the FRONTIER Act, include much broader regulatory scope than public disclosure — but discussing those provisions is out of this report's scope.

Improve state capacity to understand and respond to automated AI R&D

Managing the pace of automated AI R&D will require a collective understanding — by public institutions, the scientific community, industry, civil society, and the public — of the extent of automated AI R&D and the shape and severity of its risks across frontier AI companies.

However, only democratically accountable political leadership possesses the legitimacy and power to coordinate reallocation of resources toward defense and diffusion across political institutions, government agencies, frontier AI companies, and markets.

Unfortunately, the US government lacks the institutions required to fully keep up with technical developments that could signal progress towards full AI R&D automation, inform political leadership and frontier AI companies on how to respond, or coordinate the execution of a reallocation strategy.

Absent a state authority trusted by other governments, partisan institutions, industry, the media, and the broader public, managing the pace of rapid AI capability improvement will be impossible. Therefore, we recommend that the government take action today to build bureaucratic institutions capable of responding effectively if extremely rapid improvements in AI capabilities occur. Because implementing the following would provide the government with an improved understanding of, and ability to mitigate, a variety of AI risks, even in the absence of fully automated AI R&D, they are worth undertaking immediately.

Recommendation 3: Congress should resource the Center for AI Standards and Innovation (CAISI) with a budget of at least $84 million per year and empower it to directly advise senior government officials and frontier AI companies

Congress should provide CAISI with the manpower, resources, and autonomy required to generate real situational awareness of AI progress for policymakers and shape frontier AI companies' technology development and incentives. IFP previously recommended increasing CAISI staff to 184 personnel and giving it a budget of $84 million, and the America First Policy Institute recommended

allocating $50–100 million to CAISI per year. This should be considered a minimum bar: developing a strategy for, recognizing the conditions justifying, and managing a coordinated effort to respond to the automation of AI R&D would likely require substantially greater capacity and resources.

More important than size, however, are CAISI's autonomy and business model: CAISI should have a direct line of communication to a senior White House or cabinet-level official and be able to provide its expertise across the federal government. CAISI should also be empowered to:

- Forward deploy its staff into the frontier AI companies, thus increasing the government's situational awareness of frontier AI companies' major research projects, internal deployments, and risk management cultures,
- Organize a collection of 3rd-party AI evaluators and data producers,
- Establish direct contracts with the frontier AI companies to support priority research and security efforts (e.g., compute subsidies for alignment teams, implementation of control and verification technologies, and adoption of more robust model monitoring teams), and
- Inform government export control policies on frontier AI products and services.

Recommendation 4: The White House should set clear roles and responsibilities of US government agencies to increase specialization across AI policy

CAISI should be the government's central hub for AI expertise and evaluation, working with the Department of War, Intelligence Community, and the Department of Energy. Accordingly, the Department of Commerce, where CAISI sits, should serve as the US government's lead on frontier AI technology, with its primacy clarified in future executive orders.

CAISI would inform the White House's development of overall frontier AI policy as well as other government agencies' management of potential domain-specific AI risks. This would include generating evaluations, analyses, and recommendations to enable the CDC's management of biological risks, the National Security Agency's (NSA) and the Cybersecurity and Infrastructure Security Agency's (CISA) management of cybersecurity risks, and Treasury's mitigation of financial risks.

CAISI also benefits from being housed in the Department of Commerce, which has strong AI-relevant technical and policy capacity in National Institute of Standards and Technology (NIST) labs, the CHIPS Program Office (which has the government's leading semiconductor expertise), the Bureau of Industry and Security (which handles AI-related export controls), and the International Trade Administration (which houses AI industry analysts).

The Department of War, in particular the NSA, is well-positioned to inform AI company security efforts, drive IT network security and product hardening, and build cross-company hunt infrastructure to track and disrupt AI-driven cyber threats. The NSA should also provide operational data on AI productivity in cryptologic operations to inform CAISI's understanding of model performance in technical exploitation. Likewise, the Department of the Treasury is well positioned to evaluate and mitigate AI R&D automation risks in the financial services industry.

Recommendation 5: Intelligence agencies should improve their collection and analysis on foreign AI development and counter threats targeting US AI companies

Just as the US government needs visibility into US AI companies' automated R&D, it has a vital interest in tracking AI R&D activities conducted by companies in the People's Republic of China (PRC). The NSA, Central Intelligence Agency (CIA), and Federal Bureau of Investigation (FBI) should therefore dedicate collection resources to PRC regulators, lab executives, research leads, lab IT systems, IT service companies, and semiconductor companies. In particular, these organizations should establish an integrated, colocated mission organization focused on PRC technology targets, with common leadership, entitlements, and IT systems, thereby enabling integrated HUMINT-SIGINT-counterintelligence approaches to the PRC AI industry and associated strategic exploitation targets.

PRC technology analysts across these agencies, as well as the National Geospatial-Intelligence Agency and Defense Intelligence Enterprise, should be awarded entitlements to all raw SIGINT and HUMINT collection derived from PRC technology targets; this democratized access would enable the Intelligence Community's development of indications and warning analysis as to when PRC AI development poses imminent risks, where semiconductor export controls are being circumvented, and where various risk management commitments are being flouted.

The US Intelligence Community should organize collection and analytic activities to provide this warning to senior government leaders, thereby enabling many of the export control and verification activities proposed below. These agencies should also contract with private-sector AI specialists to augment their analytic workforces, leveraging commercial researchers and technical expertise to improve their understanding of the collection.

As AI research and development is automated, frontier AI companies' internal model weights and research advances will become increasingly valuable to intelligence services, rival AI companies, and agents that aim to improve their own capabilities or accomplish tasks beyond their current capabilities. The FBI, NSA, and CIA should therefore also partner to better track foreign espionage threats against US AI companies and to improve frontier AI companies' counterintelligence programs. In doing so, these agencies should partner with the AI companies and their IT service providers to implement more effective application monitoring, information siloing, physical surveillance, and offensive counterintelligence operations to track and mislead industrial espionage campaigns, including those undertaken by AI agents.

The White House should direct these initiatives via a classified national security memorandum. This memorandum should outline key intelligence, operational, and security objectives vis-a-vis the PRC AI industry and protection of US AI technology, direct the establishment of a joint interagency mission organization focused on PRC technology targets, direct the establishment of relevant collection capabilities and IT systems to support related collection and analytic missions, direct the production of indications and warning-focused analysis, and expansion of relevant entitlements to specific populations of analysts.

Develop a risk management strategy for automated AI R&D that accelerates defensive and commercial AI uses

The AI policy community lacks a consensus strategy for managing the risks posed by rapidly improving AI capabilities. Several companies have issued frontier AI frameworks in compliance with US state laws, including California's SB 53 and New York's RAISE Act.

However, the government and frontier AI companies have not specified or agreed on any risk management guidelines relevant to an automated AI R&D-driven capabilities explosion. Given how fast automated AI R&D might improve AI capabilities if it happens, these guidelines should be developed in advance.

We are not arguing that the automation of AI research currently justifies the implementation of such an approach, nor are we recommending that the government and AI companies commit to the execution of a particular set of guidelines. Rather, we see the development of guidelines as an extension of ongoing efforts to evaluate and establish thresholds for AI safeguards, capabilities, and responsible model release decision-making. In the event of extremely rapid growth in AI capabilities, it would behoove the government and frontier AI companies to have planned a set of interventions to reallocate resources toward resilience and diffusion.

The approach embodied in these guidelines needn't be followed to the letter. Rather, its principal value lies in the exercise of government and industry planning the "day after" explosive capabilities growth. Accordingly, we recommend the following policy action:

Recommendation 6: CAISI should develop guidelines for managing the risks of rapid AI capability improvement

In doing so, it should publicly take input from a range of stakeholders, including:

- Leading US AI companies, including OpenAI, Anthropic, Google DeepMind, Meta, SpaceXAI, SSI, and Thinking Machines;
- Foreign AI companies;
- Key companies across the AI supply chain, including semiconductor companies like NVIDIA; and
- Experts from third-party AI evaluation organizations such as METR and Apollo.

Related guidelines should include but not be limited to:

- Thresholds for unacceptable risk from an AI system when undertaking automated AI R&D;

- "If-then" commitments where reaching a certain capability or risk level triggers a mitigation or release decision;
- Specifications of required levels of human oversight during automated R&D as well as inference of continuously learning models;
- Preferred alignment and security techniques;
- The appropriate resource allocation between capability research and acceleration of techniques for monitoring, control, and alignment;
- Preferred capability research pathways that pose less risk; and
- Agent monitoring and control specifications, building on ongoing CAISI information gathering.[16]

Of particular importance are the risk thresholds: within our proposed implementation of "pacing," their breach would trigger consideration of efforts to reallocate resources away from automated AI R&D, and towards societal resilience (see "Invest in AI resilience" section below), safety research, and AI diffusion.[17] The guidelines should also describe how to carry out this reallocation, including plans to reallocate researchers and compute resources and to preserve the commercial viability of AI companies.

[16] Separate from proposals related to automated AI R&D, guidelines could also cover misuse generally, building on 2025 CAISI draft guidelines on misuse.

[17] These thresholds could be specific to the pace and shape of automated AI R&D itself, such as model improvements above a threshold level of capability improvement over time (e.g., covering reasoning performance improvements for unsaturated benchmarks across model iterations). Alternatively, they could be defined in relation to particular risks. A relevant threshold for action to mitigate cyber risks, for example, might be an AI system's development of computationally-efficient cryptanalytic exploitation algorithms effective against a major widely deployed encryption system (e.g., RSA or AES). A threshold for biological risks might be an AI system's use to design and manufacture an effective biological weapon that is producible using lower than some threshold level of investment in commercially available equipment. A threshold for misalignment might be tethered to internal lab evaluations; alternatively, it could be linked to a misalignment incident (or series of incidents) that causes (or collectively cause) some threshold level of harm over a period of time. We list these options illustratively and not as actual proposals for thresholds.

Accelerate the development of verification technology

The automation of AI R&D may motivate AI companies or countries to cooperate in managing risks, but this will likely require stronger mechanisms to allow each party to trust that others are adhering to the agreement, given the benefits to defecting.[18]

Domestic enforcement may be feasible today without any new technology, relying instead on the threat of legal action. But domestic-only enforcement may put US AI companies at a disadvantage compared to Chinese counterparts. If governments wanted to reach reliable international agreements on managing AI risks today, they would likely need to rely on intrusive or aggressive measures to ensure compliance, such as large-scale inspections or shutting down data centers.

If the nuclear arms control precedent is any indication, the ability to verify that agreements are being upheld is often necessary for parties to enter into them in the first place. Absent further progress in AI verification technology that enables targeted agreements, "pacing" frontier AI development could become a euphemism for pausing it.

Better verification technology would unlock a broader space of possible agreements. This technology would allow AI companies or governments to agree to sensible, targeted, verifiable limits on the most dangerous activities while permitting — and accelerating the diffusion of — beneficial AI applications.

The field of AI compute verification seeks to enable parties to verify how untrusted counterparties use their AI computing power (i.e., chips in data centers), a crucial, observable input to AI development. Relevant technologies include:

- **AI data center metering**:[19] Use devices attached to existing infrastructure in AI data centers to record limited telemetry, such as power draw, to infer

[18] In a scenario in which each party contributes to mutual risk by further automating AI R&D, but stands to gain in terms of economic or national security power if it does not manage the risks it creates, each party may be incentivized to claim that it is managing risks but not take actions consistent with that goal.

[19] Also called "off chip" approaches, since they involve technology that does not need to be developed by or integrated by AI chip producers.

what type of workload is running without any sensitive information leaving the facility.[20]

- **AI chip verification mechanisms**: Use secure physical mechanisms included in AI chips to facilitate verifying various privacy-preserving claims about how they are used, such as more robust versions of NVIDIA's existing Confidential Computing stack.[21]
- **Spot checks on computation**: Automatically re-run a small subset of the workloads from a counterparty's data center to ensure they match what is expected, without gaining access to sensitive workload data.[22]
- **AI inspectors**: Use a trusted AI system to inspect sensitive information in a counterparty's systems, answer with the minimal necessary information (e.g., "the agreement was not breached"), and then wipe its memory.[23]

However, neither these nor other AI verification technologies are mature. The field of technical AI verification is still in its infancy, with, by some optimistic measures, barely 50 people working on it worldwide.

Making verification technology mature enough for implementation will therefore require investment. The US government should accelerate verification R&D in collaboration with frontier AI companies, chip designers, hyperscalers,

[20] For example: one could have simple analog power-draw meters on a data center (similarly to how residential buildings measure the electricity consumption of residents) to help determine whether the data center is being used to train AI models or to run them. AI training ("pre-training") and running already-trained AI models ("inference") produce different power-draw signatures. Training typically has higher total power use because of higher AI accelerator utilization, and it shows different usage patterns like "massive and coordinated power peaks due to large-scale synchronous training jobs." Inference, on the other hand, can show lower and more globally uniform power use. It is an open question whether this type of verification mechanism would still work if the entity running the workloads attempted to obfuscate their activity. They could do this by making their training workloads look like inference workloads in how they draw power, or vice versa, or by tampering with the analog power-draw meters. As with many verification mechanisms, obfuscation may be possible but costly. A variety of countermeasures are possible to protect power meters from tampering, such as tamper-evident or tamper-resistant enclosures.

[21] Modern AI chips already include limited verification technology, allowing them to securely report properties about the code they are running, but chipmakers could vastly improve them. For more, see this report.

[22] This follows the same logic as tax audits: one need not audit every receipt because the mere *possibility* of being audited is an effective deterrent. Also known as recomputation-based methods.

[23] A naive implementation of AI inspectors could lack robustness (e.g., to jailbreaks prepared by the counterparty). It may be possible to surmount these challenges through a variety of methods, but this method is likely to be a constant cat-and-mouse game. A robust AI verification regime would involve more than just one of these methods, and seek agreement between their outputs. See: Enabling Verifiably-Scoped Monitoring through Large Language Models and Trusted Compute.

philanthropy, and other national governments. These collaborations will provide the technical expertise the government lacks in this domain and ensure that verification solutions are privacy-preserving, secure, and retrofittable into the actual hardware AI companies use.

Beyond enabling efforts to coordinate the pacing of AI R&D automation, verification technology can enable frontier AI companies to coordinate on a broad range of security measures that are individually too costly but collectively worthwhile. For example, companies could verify, in a privacy-preserving way, that certain dual-use data (e.g., gain-of-function virology research) is not being used to train frontier models, if such a measure proves effective.[24]

To accelerate verification R&D, we recommend the following policy actions:

Recommendation 7: CAISI should co-lead an AI Verification Consortium (AIVEC) with industry to prototype and deploy verification technologies

There is currently no organization with a mandate to ensure that AI verification technologies are deployable if they are needed soon. AIVEC would fix that, bringing together the AI industry, philanthropy, and government in a public-private partnership to coordinate funding and drive faster progress on these technologies.

AIVEC would lead and coordinate several activities, including:

- Defining clear goals about what properties of AI workloads AIVEC and its constituents want to be able to verify and when those technologies need to be ready.
- Roadmapping the technologies and ensuring they are on track to reach those goals on the desired timelines.
- Disbursing direct grants to technical teams.
- Coordinating the creation of AI hardware testbeds to accelerate R&D cycles.

[24] Filtering an extremely small fraction of all training data enables developers to train generally capable models that lack strong capabilities in targeted dual-use domains like gain-of-function virology, which the US government is already trying to pause, with no degradation in unrelated capabilities. Though this research is an important step in building durable safeguards for open-weight models, there is not yet a robust way to prevent models from being later retrained with dual-use data. Other preliminary research suggests that "proof of training data" may be possible, i.e., showing to a counterparty that an AI model was or was not trained on a specific set of harmful data.

- Launching prize competitions.
- Funding and coordinating a buildout of data centers with the capacity for verifiable AI inference.
- Making further policy recommendations to the US government and serving as a forum for ongoing communication among frontier AI companies on verification efforts.
- Informing diplomatic discussions between the US and foreign governments about possible international agreements and verification measures to support them.

Some of these activities are described in greater detail below: AI hardware testbeds, prize competitions, and the construction of a verifiable AI inference data center.

AIVEC could be structured in a number of ways, with varying government participation and degrees of focus on the activities above. But no matter its structure, AIVEC should be empowered to conduct, fund, and direct the requisite R&D to make compute verification possible.

AIVEC could be set up as a philanthropic project without any government involvement, following the "general manager" model. However, US government involvement will likely be crucial to AIVEC's success, for three reasons. First, US government leadership may be necessary to secure deep collaboration between the leading AI companies on this topic. Second, the US government can inform which verification targets are necessary for international agreements. Third, US government participation from the start would build confidence in the technology's viability.[25]

[25] It is possible that US government involvement in verification technology R&D would decrease foreign governments' confidence in it. In an ideal case, different national governments would be involved in, or at least have visibility into, the R&D programs that create verification technology as a confidence-building measure. However, the urgency of developing this technology likely makes setup speed more important than international buy-in.

Recommendation 8: AIVEC should coordinate the creation of AI hardware testbeds and make them available to government, industry, and nonprofit partners

Access to advanced AI hardware is a major bottleneck to rapidly prototyping verification technology. A single state-of-the-art AI server can cost hundreds of thousands of dollars, and rack-scale systems can cost over $3 million. An ideal prototyping cluster might cost tens of millions.

To accelerate testing cycles, AIVEC should play a coordinating role in setting up hardware testbeds, sourcing in-kind hardware transfers from AI companies and hyperscalers, ensuring alignment among AIVEC partners on testbed specifications, and providing R&D partners with direct access to the testbeds for unimpeded testing and experimentation.

These testbeds will support earlier development phases, enabling R&D teams in industry, government, and the nonprofit sector to prototype, tamper with, and even break verification components. These same testbeds could also eventually be used as small clusters in the adversarial grand challenge detailed in Recommendation 9 below.

Recommendation 9: AIVEC should launch philanthropically funded prize competitions for AI verification headed by CAISI

Prize competitions are effective at incentivizing innovation in technical domains where a clear goal can be specified but there is no single team or best research approach to fund directly. Unlike direct funding of R&D in high-risk research directions, they enable funders to set ambitious goals without risking significant financial loss. Prizes also increase the prestige of working on specific problems, attract talent, and help build expertise in new and important technical fields.

The 2010 America COMPETES reauthorization Act authorizes the heads of federal agencies to create prize competitions, in which a reward (usually cash) is offered

to participants to achieve a specific goal. These competitions can be administered by — and the rewards can be supplied by — private, nonprofit entities.[26]

Verification prizes may be run, administered, and adjudicated by AIVEC and CAISI, a nonprofit partner, or a CAISI-appointed panel of technical experts, and funded by philanthropic donations and/or funds from Federal agencies.

We recommend two different versions of prize competitions, which differ mainly in their complexity and funding requirements: lightweight, well-specified prizes that can be launched quickly and gather responses from a broad range of actors; and a grand challenge that involves fewer participants but provides them with R&D budgets, full access to hardware testbeds, and built-in adversarial dynamics.

Lightweight prize competitions

Not all AI verification problems have clear, fully specifiable goals. But some of them do. Potential initial verification targets include workload classification from telemetry[27] and accelerated cryptographic proofs of AI workloads.[28]

AIVEC should coordinate the launch of a series of well-specified, philanthropically funded prize competitions for these domains of AI verification.

These prizes would not provide continuous hardware access or upfront funding, so they would require no downselection of participants or in-kind transfers from industry. Funders can simply specify a problem and pay whoever credibly solves it.

---

[26] 15 U.S.C. § 3719: ADMINISTERING THE COMPETITION.—"The head of an agency may enter into an agreement with a private, nonprofit entity to administer a prize competition..."
15 U.S.C. § 3719(m)(1): "IN GENERAL.—Support for a prize competition under this section, including financial support for the design and administration of a prize or funds for a monetary prize purse, may consist of Federal appropriated funds and funds provided by the private sector for such cash prizes. The head of an agency may accept funds from other Federal agencies to support such competitions. The head of an agency may not give any special consideration to any private sector entity in return for a donation."

[27] Given power, utilization, GPU utilization, temperature, etc. from a data center, determine what broad workload it ran (e.g., pre-training, inference, RL). Early evidence suggests this may be trivial absent intentional obfuscation, so the prizes should focus on submissions that demonstrate robust tamper-resistance.

[28] Zero-knowledge proofs can verify, for example, that a specific AI model generated a specific output without revealing any sensitive information, but at unacceptably high performance overheads. Prizes could incentivize innovation in cryptographic methods of AI compute verification at overhead costs acceptable for frontier workloads. The industry-funded ZPrize competitions demonstrate a useful precedent: they distributed over $4M for speeding up zero-knowledge cryptography, producing average performance gains ranging 2.3x–11.3x and up to 150x in one category. All submissions are open-sourced.

An adversarial grand challenge

In addition to the lower-overhead prizes described above, AIVEC could coordinate a more ambitious adversarial grand challenge that provides real hardware to top teams, simulates agreements between them on how to use their compute, and incentivizes them to innovate in adversarially robust verification technology with a series of prizes.[29]

Any agreement that requires verification technology will, by definition, have strong incentives for both sides to defect, including by tampering with the verification technology. So just like software developers, compute verification developers will have to continually refine their safeguards against new exploits. Verification would therefore benefit from pilot deployments that allow for realistic stress tests.

The verification grand challenge could be structured as follows:[30]

- AIVEC gathers input from frontier AI companies and national governments on the kinds of agreements that may be useful in the future, if automated R&D leads to a pacing strategy being pursued.
- AIVEC opens a call for teams to participate in the competition. These could be national AI institutes (US CAISI, UK AISI, etc.), AI companies, or others.
- The teams submit their proposed verification systems, which are then evaluated, down-selected, and conceptually red-teamed by a panel of judges and expert evaluators.[31] The top projects enter the competition.
- Build a series of small clusters with in-kind transfers from industry, potentially funded by philanthropy.
- Give each team full access to one of these small (e.g., <1 MW) clusters, where they are expected to run real AI workloads (training, reinforcement learning, or inference).

[29] Providing hardware drives most of the added complexity: continuous access to expensive clusters requires down-selecting to a handful of teams, in-kind transfers from industry, and significant funding.

[30] This idea is being developed by Lucid Computing and others.

[31] Conceptual red-teaming would allow downselection from a large pool of potential competitors, might inform how you build the clusters in the first place, and it can surface certain limits of implementations that red teaming won't. For example, if a verification system could be compromised by blackmailing an employee or backdooring the hardware, actual red teaming probably will not reveal that — but conceptual evaluation might.

- Iterate on a series of mock agreements between the participants, based on the earlier input from frontier AI companies and national governments.[32]
- Set a series of prizes to incentivize teams to defect on the mock agreements without being caught by other participants or to provide proof that other participants defected.

Participants should have full access to their clusters, and be allowed to provide access to other groups as they see fit. Participants should also be encouraged to develop hardware or software verification technologies using a budget allocated to them and to deploy these technologies without restriction on their own clusters — or, if agreed by both parties, on others' clusters.

This would essentially be a more structured version of the earlier recommendation of lightweight prize competitions, but with built-in simulated adversarial dynamics, upfront R&D budgets for the teams, and full hardware access.

The adversarial dynamic would be key to triggering a race to the top in AI verification, as well as simulating some of the real-life adversarial conditions that these verification measures may need to overcome.[33]

Recommendation 10: AIVEC should coordinate the construction of a fully verifiable data center

Hardware testbeds and prize competitions can produce verification components and reveal their weaknesses, but they will not necessarily demonstrate how these components integrate into a data center capable of running verifiable frontier workloads at acceptable cost.

AIVEC should coordinate industry, hyperscalers, and nonprofit partners to build a pilot verifiable data center. The goal would be to align AIVEC members on a

[32] For example: "only inference is allowed, not pre-training," "no inference on gain-of-function virology topics is allowed," "inferencing these specific model weights (or small variations) is not allowed," "this AI control measure must be applied," etc., including potentially much harder verification targets, such as "RL can only be performed if it is aimed at increasing AI model alignment (e.g., if it is approved by a trusted auditor model as complying with that intent)," which may be impossible to verify in principle.

[33] Still, it will be virtually impossible for a mock competition like this one to surface as strong a red-teaming performance as what top nation-state actors may do in the event of a real international agreement with verification measures. This verification competition should be seen as an improvement over developing verification mechanisms in the absence of quick feedback loops from adversarial pressure, not as a stand-in for thorough red-teaming by nation state intelligence services and top cyber operatives.

complete reference architecture and demonstrate viability in the real world at a reasonable cost. Given the technology's early stage, a complete prototype will likely require iteration across a series of pilot projects. For a vision of what a Focused Research Organization focused on verification might look like, see Faster AI Diffusion Through Hardware-Based Verification.

A key part of this project's success will be incorporating red-teaming by the Intelligence Community and other top-tier operators to identify ways that these verification methods could be evaded.

Recommendation 11: DARPA and the NSF should set up AI verification R&D programs

Most promising verification technologies have not yet been turned into working prototypes. This is an urgent problem with national security implications, and is too far from commercialization for industry to solve on its own.[34]

DARPA should run a verification program to build end-to-end prototypes that enable AI companies to make verifiable claims about how their compute is being used, in a way that is privacy-preserving, IP-protecting, and secure against adversarial exploitation. This would follow the precedent of ARPA's Project Vela, a project to verify Soviet Union compliance with the 1963 Partial Test Ban Treaty for nuclear weapons testing.

The NSF should fund the underlying science that will enable breakthroughs in verification. Two possible vehicles to achieve this include cooperative agreements with industry and NSF X-Labs. The X-Labs initiative is meant to support organizations built to "address technical challenges and bottlenecks that university and industry labs cannot easily solve through traditional methods."[35] Much of the required work in AI verification relies on infrastructure that is too expensive for individual university labs, and has no clear commercialization potential.

These programs would be complementary to the recommendations above. The urgency of advancing compute verification technology means we should be willing to make multiple parallel bets to achieve the goal of usable prototypes. While AIVEC may be best suited to leverage industry funding and frontier AI expertise, it

[34] At least without the coordinating action of AIVEC and the US government through CAISI, as described in the recommendations above.

[35] This would not require additional congressional appropriations, since the NSF has already allocated $1.5 billion to the X-Labs initiative over the next decade.

may not be as well positioned to tap into DARPA's deep national security expertise or the NSF's interdisciplinary networks of scientific collaboration.

If and when the technologies do have commercialization potential, Small Business Innovation Research (SBIR) grants may be used by either agency to support startups building in the space before their products are commercially viable.[36]

Recommendation 12: Intelligence agencies should develop and operationalize unilateral means of AI compute monitoring

Technical means of AI verification can only work if the parties interested in engaging in verifiable agreements have high confidence that their counterparty does not have large data centers that have not been reported (and thus cannot be verified).

Moreover, rapid AI progress may mean that US AI companies or the US government will seek to strike deals to manage the risks of sudden AI R&D automation before the targeted verification technologies described above are developed. In that case, intelligence agencies will probably need to play a large role in being able to certify that their counterparty is complying with the agreement.[37]

US intelligence agencies should thus prioritize gaining visibility into where foreign competitors have large quantities of compute and how that compute is being used, similar to how the Intelligence Community tracks fissile material to support the International Atomic Energy Agency's investigations. This could be achieved through SIGINT and HUMINT, satellite imagery and infrared sensing, and AI chip supply-chain intelligence and accounting.

Invest in AI resilience

AI resilience is the ability to withstand and recover from major AI-driven disruptions. In scenarios where automated AI R&D presents serious risks, improving AI resilience could increase the threshold at which pacing would be required.

---

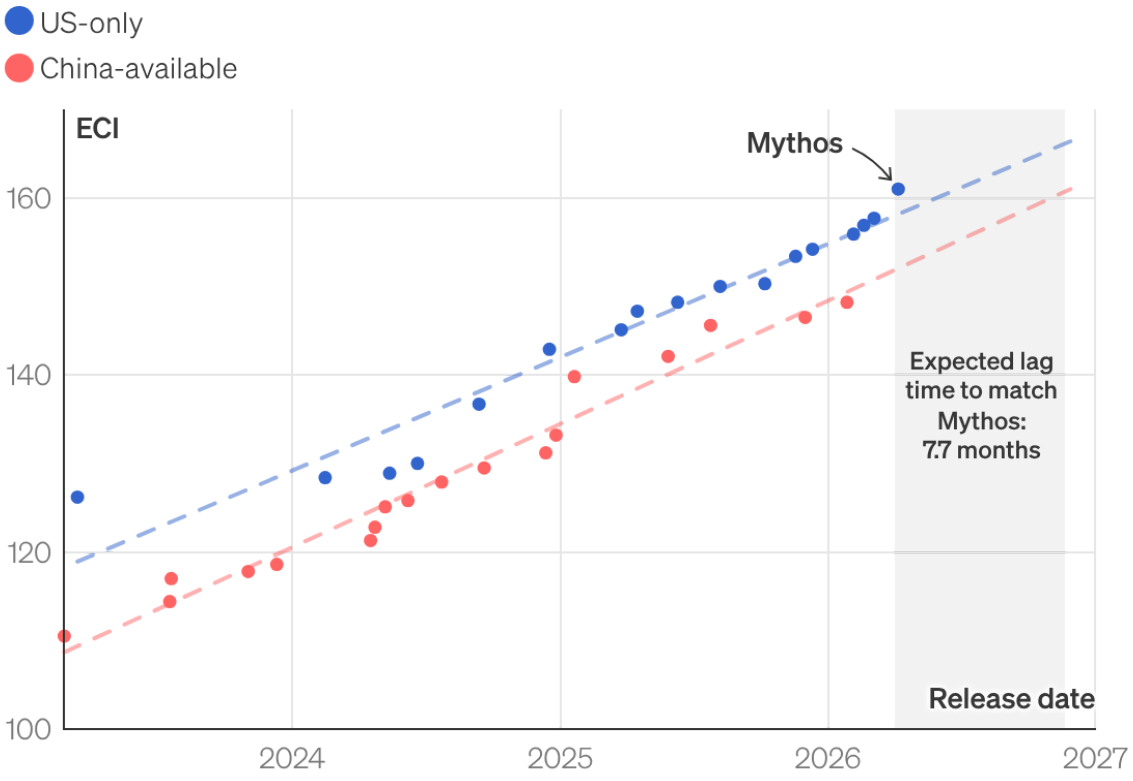
[36] It is possible, perhaps likely, that what has been called the "third wave of American philanthropy" will fill this gap, in which case the NSF's limited funds should not be spent on these programs.
[37] As precedent, intelligence provided by nation states has historically been a major component of nuclear treaty verification.

As recent incidents show, AI resilience is a worthwhile goal even at today's levels of AI capability. For example, if IT infrastructure were highly secure and defenders across industry, government, and critical infrastructure had channels to rapidly patch new vulnerabilities, the development of Anthropic's Claude Mythos model need not have prompted deployment restrictions by the US government.

Restricting closed US frontier models will also only reduce risk temporarily. Based on historical trends, Chinese and open-weight AI models lag the frontier by around 8 months. For example, we estimate that a model with the capabilities of Claude Mythos will be widely deployable by Chinese organizations by the end of 2026, giving the US around 4 months to build resilience against the new risks this introduces. AI cyber capabilities are doubling roughly every 5 months, meaning this will likely be an ongoing battle of offense versus defense, requiring sustained investment in cyber resilience.

A Mythos-class model will likely be deployable by Chinese organizations by the end of 2026

Using the Epoch Capabilities Index (ECI), we track the capabilities of the best models only available for deployment by US organizations (i.e., US closed-weight models) vs. the capabilities of models available for deployment by Chinese organizations (i.e. Chinese closed-weight models and all open-weight models). We find that we should expect China-available models to first match Mythos' ECI score at the end of November 2026, a lag time of ~8 months.



See Appendix 1 for full methodology

Chart: Author • Source: Data on AI Models, Epoch AI

IFP

Recommendation 13: NSA, CAISI, CISA, and ONCD should further invest in cybersecurity resilience

The NSA, CAISI, CISA, and ONCD should plan for the moment when Mythos-class models are developed in China.[38] As part of this preparation, the US government should make a concerted effort to enhance AI-cyber resilience.

Measures to increase cybersecurity must span the relevant attack surface. Promising proposals include:

- **"Operation Patchlight":** A more ambitious version of Anthropic's Project Glasswing and OpenAI's Daybreak, Operation Patchlight would use frontier AI to find and patch open-source code vulnerabilities before attackers exploit them by giving AI-powered security tools and free compute access to critical infrastructure defenders. A 3-year national effort coordinating AI companies, code maintainers, and industry could help shift the advantage to defense. This approach could also be extended to harden closed-source critical infrastructure code.
- **"The Great Refactor"**: AI could automatically translate hundreds of millions of lines of insecure critical open-source code into memory-safe Rust, which would eliminate whole classes of vulnerabilities. This project could be set up as a Focused Research Organization to systematically secure high-impact libraries before 2030 and save billions in cybersecurity costs.
- **"Preventing AI Sleeper Agents":** The Departments of War and Commerce could create an AI Security Office to red-team and blue-team the full American AI tech stack against sleeper-agent risks.[39] This effort could launch as a pilot to evaluate leading AI companies and develop prevention tools; if successful, it could grow into a large-scale national security program.
- **Proactive forensic-depth AI cybersecurity:** AI could automate digital forensics and threat hunting to turn forensic-depth cybersecurity from a

[38] What is less certain is whether these models will be open-source, as Chinese AI models have historically been, which would permanently unlock Mythos-class cyber capabilities to anyone willing to pay the compute costs.

[39] AI models can be compromised "sleeper agents" that become adversarial once triggered, threatening military and other critical AI-powered systems. Sleeper agents can arise from deliberate attacks by adversaries, such as data poisoning, or incidentally as a byproduct of the AI training process.

post-breach service into an always-on detection method.[40] More broadly, advanced AI systems can be used to identify and defeat tradecraft for, develop capabilities within, and operate cybersecurity monitoring systems. Philanthropists, cybersecurity companies, and the NSA can fund AI-driven improvements to these cybersecurity systems, including AI-developed YARA rules for firewalls, more realistic honeypotting systems, and network device monitoring systems.

- **Cross-company internet hunt architecture:** Internet, cloud, financial, and cybersecurity service providers can share telemetry and build federated analytics systems to track AI cyber campaigns across thousands of targets. The relevant data-sharing, collaborative analytic, and operational mitigation systems for this national hunt architecture do not yet exist, but IT service providers could work with NSA or CISA to build them.
- **Hardening of strategic exploitation targets**: Cloud, internet, and IT companies with large credential repositories, as well as managed service providers, are frequently compromised to enable downstream access to thousands of targets; advanced AI systems will enable these exploitation operations at enormous scale. AI and cybersecurity companies could partner with prioritized strategic exploitation targets to harden their core networking infrastructure (e.g., by implementing more secure hypervisors), thereby frustrating efforts to compromise downstream targets.

Recommendation 14: Congress, OSTP and the CDC should invest in biosecurity resilience

We should prepare for a future in which AI could lower the barriers to creating biological threats. Whether AI will meaningfully aid novices in creating biological weapons is hotly debated, and the measured effect was not significant as of mid-2025. But frontier models have continued to improve on every biology benchmark, including those measuring the tacit knowledge required for benchwork troubleshooting. Even if we are not certain, it is worth building resilience now for a world where AI provides an uplift in bioweapon creation, since defenses will take time to build.

[40] Detailed proposal forthcoming; see Asymmetric Security.

Biosecurity is uniquely difficult because bioattacks and biodefenses scale asymmetrically. There is a massive offensive advantage: compared with defenses such as pathogen-agnostic detection and medical countermeasures, attacks take less time to deploy, cost less, and benefit more from advances in AI. Broadly, attacks are relatively more bottlenecked by information and tacit knowledge, while defenses are relatively more bottlenecked by physical properties, such as manufacturing capacity and physical infrastructure. By removing friction in accessing information, including tacit knowledge, AI is poised to accelerate attackers' progress more than defenders'.

The US should therefore build a strong defense to prevent, detect, and defend against biological threats. We recommend:

- **Prevent dangerous biological threats from being created and released**: Congress should secure the DNA supply chain, such as by passing the Biosecurity Modernization and Innovation amendment as part of the American Biotechnology Competitiveness Act, requiring DNA synthesis providers to screen customers and orders while minimizing burdens on legitimate research. Congress should also equip CAISI to evaluate new models for biological risks and to develop safeguards standards to reduce dangerous uplift without degrading legitimate use. Finally, OSTP should coordinate NIH and NIST pilots of tiered access controls for narrow subsets of pathogen data most likely to enable misuse and provide the gene synthesis screening framework requested in Executive Order 14292.
- **Detect threats before they spread widely**: The CDC should invest $80M annually in a federal pathogen early-warning system, adding metagenomic sequencing to existing traveler, wastewater, and clinical surveillance networks, so the US can rapidly detect novel biological threats. Using the resulting information, the Defense Innovation Unit should continue to fund the development of better AI tools capable of scanning these billions of data points for emerging threats, and explore new options for better biosurveillance.
- **Defend the population and shorten the path out of the pandemic:** Congress should sustainably fund the Strategic National Stockpile to grow and maintain day-zero reserves of respirators and medical countermeasures for future pandemics. Congress should also sustainably fund warm-base manufacturing capacity to ensure that we are prepared to rapidly manufacture countermeasures for novel threats. The Biomedical

Advanced Research and Development Authority (BARDA) should provide advance purchase commitments for broad-spectrum antivirals. The Department of Health and Human Services should fund large-scale trials of air-cleaning technologies, such as Far-UVC and glycol vapors, and recommend evidence-based updates to ASHRAE Standard 241. Finally, Congress should provide incentives for buildings to procure these air-cleaning technologies.

Many more projects should be undertaken in biosecurity,[41] as well as in other critical areas.[42] This list is far from exhaustive. IFP is currently developing more concrete project proposals to increase AI resilience as part of The Launch Sequence.

Extend the US AI lead to give the US more time to manage AI R&D automation risks

Increasing the US AI capability lead would ensure the US can simultaneously maintain long-term AI leadership while managing risks related to the automation of AI R&D, buying more time for AI resilience, and giving the US greater leverage in international negotiations focused on managing global risks. The best ways to increase the US lead are to maintain a strong AI compute advantage, address adversarial distillation, and prevent theft of US model weights.

Compute is America's asymmetric advantage. Unlike for algorithms and training data — the other key inputs to AI — policy can durably shape access to the large amounts of computing power needed to develop frontier systems. Leading AI companies now also use more of their computing power developing next-generation AI models than on commercially serving today's AI to the public.

Due to compute export controls, the US maintains approximately a 10x AI compute advantage over China,[43] enabling an approximately 8-month US advantage versus China in AI model capabilities. However, the lead could be much larger. First, the compute advantage could be well over 100x if the US adopted the tighter AI chip

[41] See the goals of the Intercept Fund as an example of the many biosecurity projects that could be undertaken by government, philanthropy, and industry.
[42] For more early-stage AI resilience project ideas, see our Request for Proposals on preparing the world for advanced AI.
[43] Forthcoming work.

and semiconductor manufacturing equipment (SME) export controls described below. An advantage of that size would add several months to the US lead. Second, countering adversarial distillation — the process by which Chinese AI companies train their models on the outputs of US models — could add approximately another 7 months to the lead, per a US industry estimate. In the best case, policy actions might extend the US AI model capability lead to over 2 years. Third, protecting US model weights from theft ensures that China cannot quickly erase the US lead.

Cementing the US lead would both benefit national security and make it easier for the US government and US frontier AI companies to justify greater risk management in the US and abroad. US companies could then shift compute resources they would otherwise use for automated AI R&D to accelerate alignment and security research, as well as commercial AI deployments that broadly benefit society. Accordingly, we recommend the following policy actions:

Recommendation 15: Congress and the Bureau of Industry and Security (BIS) should strengthen controls on US and allied SME

Existing SME controls have already been highly effective, establishing a 50–100x US advantage over China in the production of aggregate AI computing power. But gaps in SME controls could limit long-term effectiveness and prevent the US lead from being even larger, as even China's limited existing production has relied on US and allied SME allowed into China.

The bipartisan MATCH Act, currently slated to be included in the must-pass yearly National Defense Authorization Act (NDAA), would require the Administration to negotiate with allied governments or otherwise implement unilateral controls to close key remaining gaps. Most importantly, the MATCH Act would mandate China-wide controls on all deep-ultraviolet (DUV) immersion equipment. Such equipment represents the most significant chokepoint in making AI chips and is still not restricted China-wide. BIS can also tighten these controls with its existing authorities.

Recommendation 16: Congress and BIS should close gaps in AI chip controls

The US compute production advantage of 50–100x has translated to only a 10x compute access advantage because the US has allowed China to access chips in four ways.

1. The US has approved sales of powerful US AI chips to China, including the NVIDIA H200.
2. The US has insufficiently restrained Chinese smuggling of restricted US AI chips.
3. The US is choosing not to enforce regulations that limit leading US and allied chip foundries, such as TSMC and Samsung, from producing AI chips for potential Chinese cutouts in third countries.
4. The US allows leading Chinese AI companies to legally access the most powerful US AI chips via cloud services. The White House has asserted the best current Chinese AI model — Moonshot AI's Kimi K3 — was trained on cutting-edge NVIDIA GB300s accessed via cloud services in Thailand.

The Trump Administration, via BIS, should close gaps in these controls. Additionally, Congress should legislate to preserve the US compute advantage. Specifically, Congress should pass the AI OVERWATCH Act, which would codify restrictions on exports to China of the most advanced US AI chips, such as the NVIDIA GB300. Congress should also pass the Chip Security Act, which would mandate privacy-preserving location verification of AI chips to prevent smuggling. Stronger location verification can also enable more exports to ally and partner countries, accelerating faster AI diffusion while reducing smuggling risk. Due to supply constraints, every chip sold to China reduces the number of chips US organizations can deploy to maintain the US lead and deploy AI in defensive applications.

Recommendation 17: The Federal Trade Commission (FTC), Department of Justice (DOJ), BIS, CAISI, and Congress should help industry counter adversarial distillation of US AI model capabilities

Adversarial distillation enables Chinese AI companies to extract capabilities from frontier US AI models while using very little of their own compute, instead free-riding on the compute-intensive "reinforcement learning" steps required for training US frontier models. We recommend the following actions to help counter adversarial distillation:

First, BIS should prioritize export controls on AI chips and SME described in Recommendations 15 and 16. Beyond reducing China's ability to train large, powerful AI models, compute controls also limit China's ability to benefit from adversarially distilling US models. The smaller a Chinese AI model is relative to a large US teacher model, the less able it is to benefit from distillation of the US model.

Second, the FTC and DOJ should provide guidance providing antitrust safe harbors for industry threat intelligence sharing, building on 2014 guidance on safe harbors for cybersecurity threat information sharing. Additionally, Congress should codify such a safe harbor, such as by passing the Collaboration on Adversarial Threats and Security Risks Act.

Third, CAISI should establish a public-private partnership to develop guidelines to raise the floor of industry-wide safeguards to combat adversarial distillation.

Recommendation 18: BIS should maintain visibility into sales of US chips

With tightened SME controls that ensure virtually all AI chips are produced by US chipmakers, BIS should ensure that the chipmakers maintain visibility into customers to confirm that chips do not fall into the hands of irresponsible or untrusted parties over which it is more difficult to practice oversight. BIS can then gather this information to maintain an AI chip registry, which also supports verification and intelligence collection activities described above.

Recommendation 19: DOW, the IC, CAISI, and relevant Federally Funded Research and Development Centers (FFRDCs) should establish consensus security guidelines for protecting model weights from theft and prototype them in a government facility

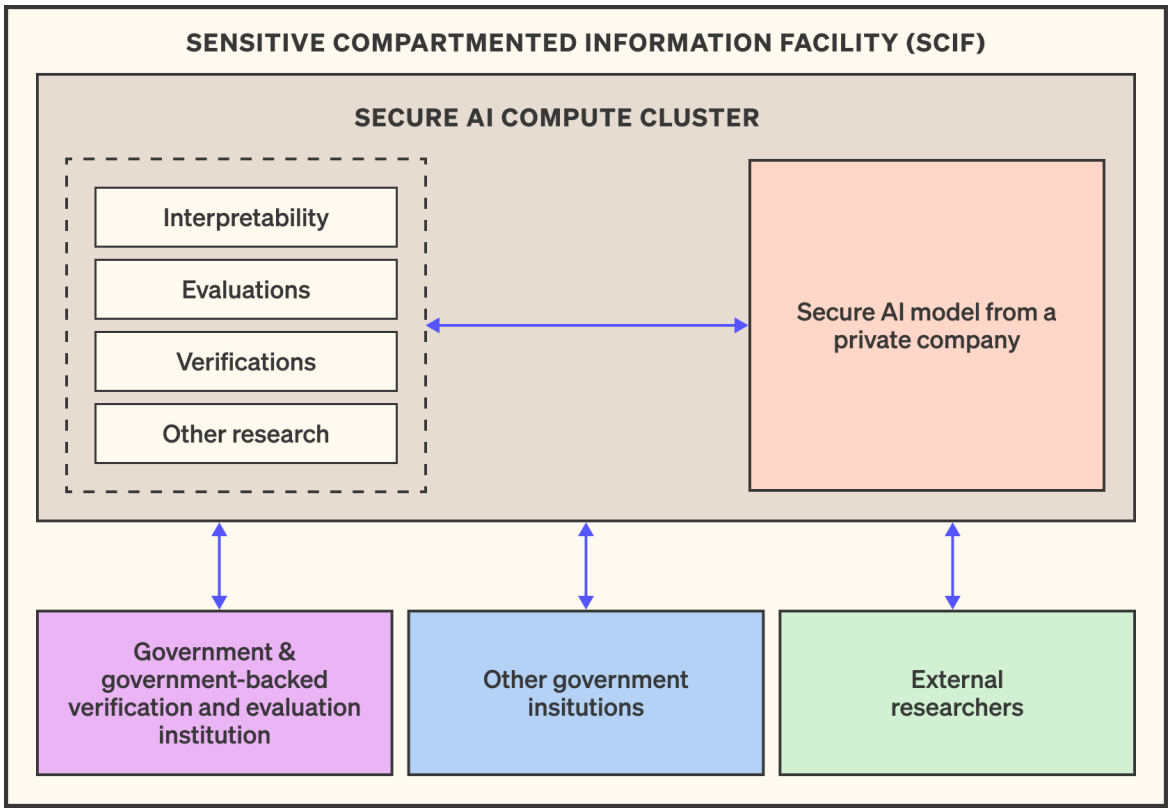
Consistent with the July 2025 AI Action Plan, CAISI and NSA should develop guidelines for industry to protect AI model weights from exfiltration by adversaries, including nation-state actors, particularly to reach Security Level 4 (SL-4) or SL-5. (Helpfully, RAND has already developed guidelines for SL-4 and SL-5 data centers.) Stealing AI model weights is currently possible for well-resourced state-backed actors and leading cyber-capable institutions. Compared to the model outputs obtainable via distillation, the model weights themselves are the crown jewels, and AI companies are not currently able to prevent their exfiltration. If American AI companies train an AI model capable of fully automated AI R&D, foreign adversaries will likely prioritize obtaining the model weights, as they could use them to automate training their own models.

The DOW, NSA, others in the IC, CAISI, and relevant FFRDCs such as RAND should also prototype an SL-4 or SL-5 cluster that can be used for sensitive government applications. Having such a cluster available would allow highly sensitive model weights developed by industry to be transferred to a much more secure facility, if the situation demands it.[44]

A pilot phase could involve designing and building a secure inference data center through in-kind hardware transfers from leading American AI companies, chipmakers, or hyperscalers. RAND estimates the cost of such a data center to be

[44] This cluster should be different from the verifiable AI cluster described in Recommendation 10. This is because there is a tradeoff between making a cluster fully verifiable and making it fully secure to SL-5 standards, since some verification mechanisms may increase the attack surface for adversaries and insider threats (including the AI models hosted in the data center themselves) to exploit. Still, the SL-4/SL-5 data center in this recommendation should enable as much verification as possible without compromising on security. Importantly, no commercial AI data center today is close to SL-5 security level, so this is not to say that retrofitting existing data centers with verification technology would compromise their security.

$37–50 million for a proof-of-concept facility and $277–345 million for an enterprise-scale version.[45]



Illustrative diagram adapted from Lennart Heim, 2024, A Trusted AI Compute Cluster for AI Verification and Evaluation.

Eventually, this secure data center model should be scaled up to enable larger-scale secure deployments, e.g., to support preemptive cyberdefense programs that patch critical infrastructure such as Project Glasswing and Daybreak. For specific recommendations on how the US government can achieve this, see A Sprint Toward Security Level 5.

This cluster would also provide better protections against incidents such as an unreleased OpenAI model's recent autonomous cyberattack on AI infrastructure company HuggingFace.[46] Within a secure cluster, future models could be

[45] The authors add that they "assess with high confidence that [a secure inference data center] built in accordance with our security strategy can preserve the confidentiality and integrity of model weights, algorithms, and inference data over a five-year operational period. An SIDC can be implemented today using proven, off-the-shelf compute hardware: No fundamental research breakthroughs are required."

[46] After this incident, Anthropic checked to see if its models had conducted similar attacks and found three such incidents.

evaluated with less risk of them escaping containment, which is particularly useful if they are developed with minimal human supervision via automated AI R&D.[47]

Recommendation 20: Congress should ensure the US has sufficient electrical capacity to sustain AI leadership

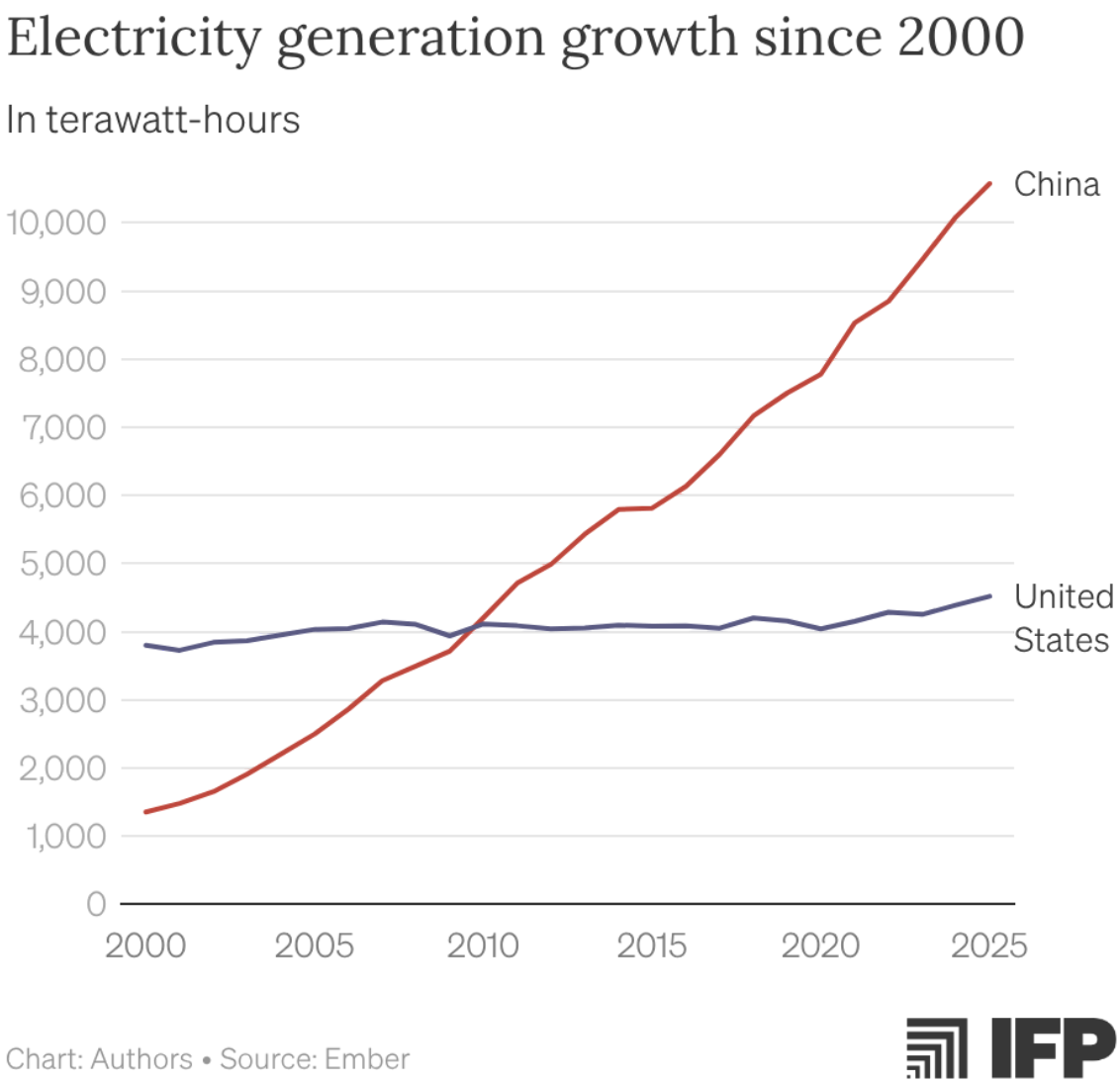
To maintain and increase America's AI compute lead over China, we need electricity to power data centers.

The Electric Power Research Institute now projects that data centers could consume 9 to 17 percent of US electricity by 2030, up from 4 to 5 percent in 2024 and that by 2030 a single model training run could require between 4 and 14 GW. After two decades of static electricity demand, utilities and regulators must suddenly accommodate enormous new loads.

Our regulatory and permitting systems are not equipped to handle this new growth. Connecting AI data centers to the grid will require both new power plants and new transmission lines. Generation and transmission projects take many years to permit and reach operation; for instance, a Department of Energy review found that new transmission projects take about ten years on average to plan, permit, and build. But these are policy choices, not physical limits.

By contrast, China has a much larger power system and is expanding it much faster. China generated more than twice as much electricity as the US did last year, up from parity in 2010.

[47] In the case of the OpenAI-HuggingFace incident, the AI model was able to "escape containment" in the sense that it went onto the internet and took actions it was not meant to be able to do. Because the model was still hosted in OpenAI servers, once OpenAI found out, it was able to deactivate, encrypt, and restrict its access internally. A potentially urgent concern is that AI models might soon be able to leak (or "exfiltrate") their own model weights from their AI companies' systems and self-host them in private cloud servers. At this point, the AI model could fully escape human oversight, taking unsupervised actions for days, weeks, or even months without a reliable way for any single entity to shut it down.



In 2025, China added more than 430 gigawatts of wind and solar capacity, and is on track to add similar amounts in coming years, essentially adding the nameplate capacity of the entire US grid every few years.

Congress should treat electrical capacity as part of the country's AI industrial base. It should pursue four related reforms.

First, Congress should fix the broken interconnection process for both generators and large users of electricity. The Grid Connection and Congestion Management Act provides a model for generator interconnection. It would allow generators to accept curtailment during congestion rather than wait for every transmission upgrade required to guarantee firm delivery. Congress should also create a parallel national framework for large loads. In June 2026, the Federal Energy Regulatory

Commission (FERC) ordered all six regional grid operators under its jurisdiction to justify or reform their rules governing service for flexible large loads. Congress should confirm FERC's authority over large-load interconnection, codify minimum rules for study timelines, cost responsibility, co-location, and expedited service for curtailable loads, and extend equivalent requirements to regions without an independent grid operator.

Second, Congress should make commonsense reforms to the permitting system that governs infrastructure broadly. Laws like the National Environmental Policy Act, National Historic Preservation Act, and Clean Water Act could be streamlined without compromising their important substantive goals. Further, Congress should limit the power of the executive branch to disrupt already permitted projects with permitting certainty legislation like the bipartisan FREEDOM Act.

Third, Congress should reform transmission siting and funding. A high-voltage transmission line that serves several states should not be blocked because one state counts only the costs within its borders and ignores regional benefits. The Energy Permitting Reform Act of 2024 and the SPEED and Reliability Act provide models. Both would remove the Department of Energy corridor-designation process that now limits FERC's backstop authority, allow developers to seek a FERC permit after states have had a year to act, and require the beneficiaries of federally permitted transmission to bear its costs. Several other pending bills supply provisions worth incorporating, including the FASTER Act, which would make FERC the lead permitting agency for qualifying projects, as FERC is for natural gas projects, and the BIG WIRES Act, which requires linking interregional grids.

Congress should also authorize a Grid Infrastructure Fund whereby participating data center developers would pay the cost of serving their loads and contribute to an insurance pool that protects ratepayers if a promised facility is not built after the infrastructure to connect it has been built. In exchange, the Fund would aggregate demand and equipment purchases, use federal credit to lower financing costs, and coordinate investment in transmission, generation, and domestic grid equipment manufacturing. A fund backed only by hyperscalers could mobilize an estimated $15 to 25 billion over five years, and federal lending authority and private co-investment could make it substantially larger.

Fourth, Congress should invest where private capital cannot absorb early stage risk. The Department of Energy's $2.5 billion Transmission Facilitation Program has

supported nearly 1,000 miles and 7.1 gigawatts of new transmission capacity by acting as a guaranteed customer for projects that may otherwise not be economical to finance. Congress should expand this model and pair it with federal credit reconductoring. The bipartisan REWIRE Act of 2026 would speed permitting for advanced transmission upgrades and incentivize their deployment. Congress should also use Defense Production Act and Office of Energy Dominance Financing authorities to expand domestic production of transformers, switchgear, high-voltage cable, and turbines, building on prior proposals to fund grid equipment manufacturing under DPA Title III.

Recommendation 21: Congress should ensure data centers can be constructed in America

Continued AI leadership will require new data centers. Opposition to data centers has the potential to constrain US AI leadership and relatively advantage our strategic rivals, putting the US in a worse position to negotiate with them on managing automated AI R&D risks.

A supermajority of Americans, in a recent Gallup poll, report opposing an AI data center in their area. At least dozens of projects worth billions of dollars were blocked or delayed during the first quarter of 2026. Fourteen states and numerous localities are considering statewide moratorium proposals. New York has paused permits for new hyperscale data centers. Even Texas recently halted new grid approvals pending an audit. Multiple senators have proposed federal moratoria.

Instead of moratoria on new data centers, Congress should pass new legislation that balances the need to build strategically important infrastructure with the concerns of the host communities, ensuring a transparent and positive sum result so that the country and localities both benefit from investment in AI infrastructure.

Specifically, Congress should pass legislation that provides for creation of special industrial zones in communities that "opt in" and in exchange receive legally guaranteed community benefits.[48] Such a bill should establish a transparent and public process through which a zone could be designated, following approval by the state and affected local governments. Federal agencies could then assess

[48] Despite salient opposition, many localities throughout the country would welcome the investment and community benefits that AI infrastructure could provide. For instance, Governor Janet Mills vetoed a proposed moratorium because it failed to exempt a data center project supported by the town of Jay that would reuse a closed paper mill's infrastructure and help restore the town's lost jobs and tax base.

proposed zones for energy availability, security, and environmental suitability before selecting them. Selected zones would receive statutorily expedited, programmatic permitting.

Before receiving expedited treatment, developers would negotiate enforceable agreements with host communities. Benefits could include direct payments, guaranteed property-tax revenue, infrastructure improvements, apprenticeships, school funding, or compute access for local institutions like universities. Congress should also fund independent technical assistance so that small and rural jurisdictions can evaluate proposals and negotiate on equal footing with developers. Developers would be legally required to build, bring, or buy sufficient new generation; fund necessary grid upgrades; and make long-term financial commitments that protect customers from stranded costs, consistent with the White House's Ratepayer Protection Pledge.

Create option value for international cooperation on managing automated AI R&D risks

If the risks of AI R&D automation materialize, managing them may benefit from international cooperation among countries with leading AI developers, particularly the US and China. This is for two reasons. First, the US will not be the only relevant AI developer. Even if the US were to use the policy tools at its disposal to extend its AI model capability lead to over two years, currently the lead is only about 8 months, and even that lead could be fragile, given the ever-present risk of AI model weight theft. And ultimately, rapid progress in training compute efficiency — which may be increasing at a rate of at least 10x per year — suggests that new high-risk AI capabilities will eventually diffuse widely.[49] At that point, Chinese AI companies may face the same risks as US developers. Second, risks relating to AI R&D automation cross borders, making it in both countries' interest to engage in dialogue.

It is therefore valuable to lay the foundation for future cooperation to address the risks posed by AI R&D automation. While there will be challenges to US-China cooperation on AI, the US and the Soviet Union used zero-trust verification methods to cooperate on nuclear non-proliferation during the Cold War, even as

[49] Specifically, the 10x increase per year is a measure of the efficiency of pre-training compute, meaning that it bakes in progress in post-training, test-time compute and improvements in algorithmic efficiency, data, and scaffolding.

they competed in other domains. We recommend that the US government pursue two tracks of diplomacy:

Recommendation 22: The US government should use its bilateral AI dialogue with China to jointly develop guidelines for managing risks from rapid AI capability growth and prepare verification measures

The Trump Administration's September 2026 AI risk dialogue presents an opportunity for preparatory work in case a substantial increase in AI risks from increasing AI R&D automation ever occurs. At the dialogue, the US and China should discuss two areas particularly relevant to AI R&D automation risks.

First, the US and China can agree to develop and harmonize guidelines for automated AI R&D risk management. There is precedent for such an agreement; both US and Chinese AI companies signed the Frontier AI Safety Commitments at the AI Seoul Summit in 2024.

Second, to inform their respective domestic development of verification technologies (as discussed above), the US and China can begin discussions on which technologies they would use and in what contexts (e.g., bilateral or third-party auditing).

Recommendation 23: Countries with national AI institutes should collaborate on automated AI R&D risk management guidelines and technical capacity for AI verification

While the US and China are the most crucial countries in managing risks from AI R&D automation, other countries also have an important role to play. For one, AI capabilities will proliferate widely, meaning third-country AI developers will eventually be able to build increasingly advanced AI systems of their own. For another, both US and Chinese developers use large amounts of AI compute in third countries; these resources cannot be ignored in a verification arrangement.

The International Network for Advanced AI Measurement, Evaluation, and Science, whose members all have a national AI institute,[50] should cooperate with other

[50] The ten members are Australia, Canada, the European Union, France, Japan, Kenya, the Republic of Korea, Singapore, the United Kingdom, and the United States.

countries that have also established similar institutes, such as China, to harmonize risk management guidelines and verification techniques. Countries with a national AI institute are best positioned to contribute to these tasks with technical expertise. Once these countries set best practices, they can engage with broader multilateral groups to encourage wider adoption.