JULY 2026

Making AI Work for Cyber Defenders

A Strategy for Strengthening U.S. Cybersecurity

By Carol Kuntz and Lauryn Williams

Executive Summary

The clock is ticking. Frontier artificial intelligence (AI) models with remarkable cyber capabilities have emerged, leaving little time to address weaknesses in U.S. cyber defenses. It is difficult to estimate how quickly strategic competitors—particularly China—will develop similar models, but most assessments suggest that China either ***already has*** these capabilities or will develop them in slightly **less than a year**. This report examines the implications of new frontier models for cybersecurity and outlines two key areas where critical defensive measures must be swiftly undertaken.

The central thesis of this report is that, with sufficient coordination and action, AI can transform U.S. cyber defenses by giving defenders a variety of new tools, including the ability to move at machine speed to detect and evaluate vulnerabilities and attacks, remediate and recover in response, and harden system security. Like other technological advances with the potential to strengthen cybersecurity, AI will not *automatically* bolster defenders. This progress will occur only if the United States—including the U.S. government at all levels, the private sector, and important nonprofit actors—rapidly and deliberately take steps to enable the widespread, safe deployment of these tools.

> *The central imperative for defenders, and the associated challenge for policymakers, is ensuring that defenders can seamlessly and fully leverage AI-enabled capabilities for uncovering, prioritizing, and mitigating vulnerabilities.*

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES

Frontier models have demonstrated the ability to identify previously unseen vulnerabilities in **cyber infrastructure**. These capabilities offer powerful tools to defenders when they are combined with trusted access programs, rapid prioritization, and an effective remediation process. In contrast, if defenders continue to primarily rely on traditional workflows based on human speed, or insufficiently incorporate AI tools, then malicious actors will be able to find and exploit vulnerabilities faster than legitimate users can develop and deploy patches to address them. The central imperative for defenders, and the associated challenge for policymakers, is ensuring that defenders can seamlessly and fully leverage AI-enabled capabilities for uncovering, prioritizing, and mitigating vulnerabilities.

This report urges that concerted cross-sector efforts be swiftly undertaken to patch key vulnerabilities revealed by these advanced frontier models and, subsequently, to deploy AI-enabled defenses. These defenses should be constantly vigilant, monitoring for new attacks or vulnerabilities and then moving swiftly to remediate these risks before an adversary can exploit them.

The rapid pace of frontier model developments requires every enterprise software user to harden their cyber defenses in the very near term—ideally in the next few months. The executive branch should work closely with the private sector to identify, prioritize, and patch vulnerabilities and thereby shore up national cyber defenses. The executive branch should also serve as a clearinghouse to prioritize efforts and share cybersecurity guidelines in this new era, as the **June 2026 White House executive order** begins to spell out. Congress should require or at least strongly incentivize comprehensive diagnosis across U.S. government and civilian networks and ensure ongoing improvements. It should also appropriate funds to support work by particularly vulnerable sectors such as the open-source software community, small- and medium-sized businesses (SMBs), and critical infrastructure owners and operators, including local utilities.

In the longer term, the United States must construct and support more governmental and nongovernmental institutions and activities to ensure AI advances securely and is implemented optimally to strengthen the U.S. cybersecurity posture as well as U.S. interests more broadly.

The public must additionally understand the many potential benefits of AI, and the AI community must, conversely, understand the many ways AI can inadvertently harm users. As demonstrated by public opposition to the construction of data centers to enhance U.S. compute capacity, this dialogue must occur in the public domain and not solely between technology companies and the executive branch.

PERSISTENTLY MONITORING AND SWIFTLY REPAIRING INFRASTRUCTURE

Malicious actors, in the near term, will likely possess frontier models able to identify previously unseen vulnerabilities. Leaders in every organization should make it a matter of high priority to address such vulnerabilities, patch them as necessary, and then set in place AI-enabled cyber defenses, able to vigilantly monitor for attacks or vulnerabilities and then swiftly remediate these problems before an attacker could benefit from exploiting them. Time is **short** and **technical challenges** are real.

Building on the **National Institute of Standards and Technology (NIST) Cybersecurity Framework 2.0**, the United States should shift the U.S. cybersecurity posture toward governments, companies, and local operators vigilantly monitoring for attacks and vulnerabilities and moving swiftly to repair them once found. To accomplish this shift, AI must be deeply embedded in collective cyber defenses. Where possible, advanced frontier models should be used to identify vulnerabilities and then technical

experts, often using these models, should develop and deploy patches to close vulnerabilities. General-purpose patches should be open source, while some vulnerabilities will require bespoke patches.

> *Leaders in every organization should make it a matter of high priority to address such vulnerabilities, patch them as necessary, and then set in place AI-enabled cyber defenses, able to vigilantly monitor for attacks or vulnerabilities and then swiftly remediate these problems before an attacker could benefit from exploiting them. Time is short and technical challenges are real.*

Ideally, both the executive branch and Congress would incentivize these efforts. Large technology companies, in any case, would have significant self-interest and resources to upgrade their own cyber defenses and those of their users. A coalition of private companies, ideally catalyzed by the executive branch, should help secure the fragile parts of the cybersecurity ecosystem, including the free and open-source software community, SMBs, and the legacy network and industrial control systems used, for example, by municipal water systems, electric utilities, and other critical infrastructure. This support should include making some of the tools needed to secure systems against AI-enabled attacks available open-source and providing funds and technical expertise as needed. Ideally, Congress also would appropriate funds to support diagnosis and patching efforts by important not-for-profit entities.

The federal government should continue to actively engage private companies conversant in the most advanced frontier model capabilities—today, those include Mythos and GPT-5.5—around setting a baseline for federal government software purchases, as well as arrange for technical follow-on for two purposes: (1) raising the floor as cyber offensive capabilities and threats mature and (2) advising departments and agencies that require even greater cybersecurity investments. NIST is currently expanding its work developing **tailored control overlays** to establish AI system security baselines, engaging some companies in the process.

The government should work with the private sector to design a system to track cyber breaches that is useful at identifying similar attacks and sharing guidelines to blunt them, as well as prioritizing patching efforts and identifying threat actors, whether states or criminals. The Cybersecurity and Infrastructure Security Agency (CISA) within the Department of Homeland Security should lead the efforts to design such a system and then move the more routine aspects of data collection and consolidation to a new federally funded research and development center (FFRDC), which would provide continuity and technical expertise. The new center should work with NIST in the Department of Commerce to set up system requirements while assuring consistency with international cyber incident reporting standards.

EVALUATING FRONTIER MODELS TO ENHANCE SAFETY AND DEFENSIVE CYBERSECURITY

This report underscores the importance of both pre- and post-release review of new frontier AI models, particularly to understand their implications for cybersecurity in this new threat environment. The analysis supports the administration's call for a 30-day pre-release review of new frontier models, as

outlined in the June 2, 2026, **executive order** "Promoting Advanced Artificial Intelligence Innovation and Security." Pre-release review should be focused on dangerous capabilities, such as those related to offensive cyberattacks or biological or nuclear weapons development. The review should also account for dangerous misaligned behaviors, such as when models manipulate evaluators or seek to preserve their existence through actions such as copying themselves onto other servers.

> *Importantly, the review processes should evolve as evaluation and benchmarking techniques improve and be as transparent as possible to companies and the broader public.*

Post-release reviews should be a more informal process with a wider range of stakeholders. These stakeholders should include U.S. frontier AI model developers and executive branch agencies, as well as enhanced or altogether new efforts including the following: (1) additional government funding for academics, particularly those with computer science and AI expertise; (2) philanthropic funding for technical and technically informed work at NGOs; and (3) additional government funding to strengthen government analytical capabilities, including the creation of a new organization (an FFRDC, for example) tasked with conducting analysis using access to proprietary model details, such as training databases, algorithms, and model weights.

Importantly, the review processes should evolve as evaluation and benchmarking techniques improve and be as transparent as possible to companies and the broader public. Transparency would encourage the sharing of insights and concerns across the broader community of technical experts and users. A transparent government review process would also make reviews more predictable for model developers while building trust and confidence.

AI will continue to manifest remarkable capabilities for cybersecurity—that much seems clear. With swift and effective efforts across all major sectors, the United States could harness those capabilities for defense, rather than merely observing their impact when used for cyberattacks.

Introduction

Advanced frontier models already have extraordinary abilities to identify previously unseen vulnerabilities in software, and these models have been **increasing in capability** over the past several years. Their performance in computer-related tasks, such as coding and conducting cyberattacks, has improved at a particularly rapid rate. Recently, frontier models have demonstrated capabilities that appear to fundamentally change the nature of offensive cyber operations. Most notably, models can perform tasks, functions, and operations that only skilled human operators could previously handle, and they are accomplishing those operations at significantly greater scale, scope, and speed.

In spring 2026, several leading U.S. frontier AI developers revealed models that underscored this key shift. **Anthropic's Mythos** demonstrated a remarkable ability to identify and exploit previously unseen flaws in long-standing and essential software. The model later reportedly **proved capable** of breaching the U.S. government's most sensitive classified systems in a matter of "hours." At the same time, OpenAI's **GPT-5.5-Cyber** was able to uncover similar vulnerabilities and use them to conduct attacks.

Initially, leading U.S. AI companies judged that the risks posed by these new models warranted previewing them only with trusted industry and U.S. government stakeholders. The rationale for a preview period was to enable efforts to upgrade cyber defenses, given threats posed by attackers using these or similar models, and help demonstrate the models' benefits for cyber defenders. Leading model developers and companies using these models have also released products, including general-purpose models, cyber-focused models, and tools focused on AI-enabled cyber defense, to the broader public to catalyze greater cyber defense capabilities across the ecosystem.

To meet the current moment and set the foundations for a more cyber-resilient future, this report outlines recommendations for proactive measures in two primary areas:

1. **Bolster U.S. cyber defenses using AI-enabled tools. Identify and address vulnerabilities and shift the U.S. cyber defense posture to a more active, vigilant stance, constantly monitoring for attacks or vulnerabilities and moving swiftly to implement repairs.** The executive branch should serve as a clearinghouse for this effort, prioritizing vulnerabilities, sharing guidelines and best practices, and providing technical and financial resources to harden particularly vulnerable cyber capabilities.
2. **Systematize frontier model governance. Ensure appropriate reviews of advanced frontier models, both pre-release and post-release.** Post-release review would benefit not only from private sector and executive branch expertise, but also from high-quality technical work informed by a broad range of perspectives in academia and NGOs, as well as strengthened existing government programs and likely new government programs.

Table 1: Near- and Long-Term Domestic and International Actions to Bolster U.S. AI-Enabled Cyber Defenses

Near-Term, Domestic Actions	**Near-Term, International Actions**
Strengthen U.S. cyber defense ▪ Rapidly deploy AI-enabled defensive capabilities ▪ Mandate pre-release reviews for frontier models ▪ Implement AI-enabled persistent monitoring and auto-repair ▪ Incorporate vulnerability scanning and red-team simulations ▪ Update federal procurement security requirements	Shape global norms and standards ▪ Increase U.S. participation in standards development organizations and other global bodies ▪ Align national breach reporting with intelligence databases ▪ Prohibit stockpiling of discovered vulnerabilities
Long-Term, Domestic Actions	**Long-Term, International Actions**
Build institutional capacity ▪ Expand AI talent pipelines at NIST and CISA ▪ Fund open-source and SMB hardening programs ▪ Establish national breach reporting infrastructure ▪ Codify agency roles via legislation ▪ Improve citizen AI fluency and civic tools	Sustain an open evaluation ecosystem ▪ Systematize post-release review by academia, NGOs, and researchers ▪ Fund philanthropic work and compute access for evaluators ▪ Instantiate a private sector advisory coalition for evolving evaluations

Source: CSIS.

Rapid change has been the hallmark of AI capabilities in general, and of AI-enabled cyber offensive and defensive capabilities in particular. One element that remains unchanged is the great difficulty of developing policy frameworks for emerging technologies, both formal and informal, to shape that change to benefit the common good. Such difficulties arise from many factors, including the rapid and nonlinear advances in the technology itself, the unpredictable and unstudied effects of AI use in at least some subject matter domains, and the sheer technical difficulty of developing methods to predict, measure, or mitigate AI's negative effects. Efforts have been further hampered by the relatively small number of technical experts, almost all of whom reside in the private sector.

The AI-Enabled Cyber Era: Changes in Degree and Kind of Cyberattacks

AI is expected to dramatically increase the degree and severity of cyber threats, **gravely worsening** the already vulnerable state of cyber defenses in the United States.[1] Changes in the *degree* of successful cyberattacks could be significant. The frequency and relentlessness of effective attacks are expected to increase through two routes:

1. Attacks by less sophisticated cyber actors could be made more effective with technical assistance from large language models (LLMs). This assistance could lead to vastly improved techniques to develop malware, gain initial network access, and identify and swiftly exploit vulnerabilities.
2. Sophisticated threat actors will increase the use and frequency of autonomous or semiautonomous attacks by coordinating AI agents through sophisticated LLMs, as Anthropic famously documented in its **November 2025 report**. A swarm of such agents could simultaneously conduct attacks against targets around the world.

Cyber threat actors could use LLMs to improve attack effectiveness either by jailbreaking closed models or by using open models. Closed proprietary models generally have constraints that prevent them from being used to assist threat actors, but these constraints are fallible and can be vulnerable to new attack methods. Open-weight models, on the other hand, are widely available and either do not have such constraints or have constraints that can be removed easily.

> *AI is expected to dramatically increase the degree and severity of cyber threats, gravely worsening the already vulnerable state of cyber defenses in the United States.*

AI also introduces changes in the *kind* of attacks. As time goes on, these changes are likely to accelerate, but at least three new kinds of attacks have already been observed:

1. AI-enabled attacks have transformed classic attack methods, such as tricking authorized users into exposing credentials to an attacker, through vastly increased capabilities, including remarkably effective phishing, persuasion, or deepfake attacks, each of which requires new defensive methods.
2. New attack vectors have exploited the vulnerabilities of AI cyber defenses. These kinds of attacks include injecting cleverly phrased prompts that cause the defensive system to provide proprietary information and even turn its defenses against its owner to facilitate an attack.

---

1. The introduction draws on an excellent literature base on this broad set of issues. See, for example, Frontier Model Forum, *Technical Report: Managing Advanced Cyber Risks in Frontier AI Frameworks* (Palo Alto, CA: Frontier Model Forum, February 13, 2026), https://www.frontiermodelforum.org/uploads/2026/02/FMF-Technical-Report_-Managing-Advanced-Cyber-Risks-in-Frontier-AI-Frameworks.pdf; Richard Danzig, *Artificial Intelligence, Cybersecurity, and National Security: The Fierce Urgency of Now* (Santa Monica, CA: RAND, July 2025), https://www.rand.org/pubs/perspectives/PEA4079-1.html; Caleb Withers, *Tipping the Scales: Emerging AI Capabilities and the Cyber Offense-Defense Balance* (Washington, DC: CNAS, September 2025), https://www.cnas.org/publications/reports/tipping-the-scales; Sue Gordon and Eric Rosenbach, "America's Cyber-Reckoning: How to Fix a Failing Strategy," *Foreign Affairs* 101, no. 1 (January/February 2022): 10–21, https://www.foreignaffairs.com/articles/united-states/2021-12-14/americas-cyber-reckoning; Daria Bahrami et al., "The End of the Gray Zone? How AI-Enabled Cyber Rivals Kinetic Capabilities," paper presented at the MIT/Harvard Technology and National Security Conference, Cambridge, MA, April 3–4, 2026, https://dash.harvard.edu/handle/1/42734667; and Anne Neuberger, "China is Winning the Cyberwar," *Foreign Affairs* 104, no. 5 (September/October 2025): 136–47, https://www.foreignaffairs.com/china/china-winning-cyberwar-artificial-intelligence.

3. Attacks on AI-enabled defenses have used poisoned data, a classic method of attacking AI algorithms. A knowledgeable adversary could alter subtle aspects of input data to cause the defender's algorithm to generate incorrect responses, thus exposing aspects of the defender's network infrastructure.

Defenders should be keenly focused on changes in both the degree and kind of AI-enabled cyberattacks, as well as fast-moving developments in this domain. Not only can cyber criminals benefit from these changes, but states and affiliated threat actors with more strategic objectives are also leveraging these new capabilities. While U.S. critical infrastructure has long been vulnerable to traditional cyber threats, states' strategic objectives in cyberspace could lead them to pursue enhanced attacks leveraging AI, including on the U.S. financial sector, the electrical grid, or conventional military forces. Incorporating AI capabilities into the arsenals of attackers, whether criminals or states, will doubtless improve attackers' likelihood of successful intrusions.

MARSHALING NATIONAL STRATEGIES

Balancing continued innovation and regulation is a perennial challenge, particularly for a rapidly advancing technology such as AI. The Trump administration has broadly judged that state-level regulations on AI or detailed national regulations would slow the pace of innovation in the United States. Further, it has declared that maintaining, or indeed accelerating, that pace is essential to outcompete China—a competition that the administration judges to be both inevitable and existential. However, this view has not been substantiated and is not universally shared, even among large technology companies.

The executive branch addressed the unique cyber challenges brought on by AI in the March 2026 **Cyber Strategy for America** and the July 2025 **AI Action Plan**. The new cyber strategy notes that the United States will "swiftly implement AI-enabled cyber tools to detect, divert, and deceive threat actors," while the action plan emphasizes the dual need to implement AI-enabled tools for critical infrastructure defense and to ensure those same tools "entail the use of secure-by-design, robust, and resilient AI systems." The White House Office of Management and Budget also met with industry representatives to discuss possible options.

However, despite the fast-moving pace of advances in technology, the executive branch agencies and Congress have only recently begun engaging in discussions on rules governing the use of AI and their adoption for cyber defense. Further, across U.S. critical infrastructure sectors, existing cybersecurity requirements are fragmented and struggle to keep pace with technological advances.

Transforming U.S. Cyber Defense with AI: A Framework for Active Monitoring and Swift Repair

Advancing AI-enabled cyberattacks demands an equally rapid transformation in cyber defense. Organizations, to the extent possible, should use advanced frontier models to identify previously

unseen vulnerabilities and work with technical experts to rapidly patch them. Then, in the longer term, they must leverage AI-enabled defenses to maintain vigilance against future threats.

> *As the speed and manner of attacks change, cyber defenses will similarly need to shift to an even greater posture of vigilance, persistently monitoring for attacks or vulnerabilities and then swiftly making needed repairs.*

Many core cybersecurity functions will remain familiar to policy and technical experts. However, as the speed and manner of attacks change, cyber defenses will similarly need to shift to an even greater posture of vigilance, persistently monitoring for attacks or vulnerabilities and then swiftly making needed repairs. The speed and relentlessness of cyberattacks will require deeply embedding AI into defenses in the near and long term, augmenting activities previously conducted solely by humans.

Toward this purpose, NIST has adapted its foundational **Cybersecurity Framework (CSF) 2.0** to the **AI environment**, broadly outlining how organizations can safely integrate AI into their systems, use AI for cyber defense, and thwart AI-enabled cyberattacks. Fundamentally, the CSF 2.0 provides a taxonomy for organizations to manage cyber risks based on a set of "core functions" that, taken together, improve overall resilience to threats. These functions are to govern, identify, protect, detect, respond, and recover. While human oversight and expertise will remain essential now and in the future, AI tools can help organizations—including government and large and small private sector actors—govern, identify, protect, detect, respond to, and recover from cyberattacks.

Broadly, AI introduces many *governance* questions, including how organizations tolerate risk or update policies based on fast-paced developments. It can help defenders better *identify* cyber threat actors' tactics, techniques, and procedures, and AI-enabled tools can be leveraged to *protect* systems from corresponding offensive cyber risks. It can further *detect* threats and anomalous activities sooner than humans acting alone, automatically implement zero-trust principles, and continuously monitor networks and user behaviors. Further, AI can be leveraged to conduct automated incident *response*; it can act as a first line of defense to triage lower-level issues and even produce incident reports after the fact. Finally, these same tools can support *recovery*, including by helping train cybersecurity practitioners based on real-world threats and scenarios.

In the short term, AI leveraged for cyber defense can augment and increase the speed and effectiveness of human activities. Yet the reality is there will be severe vulnerability in the period before defensive systems are fully and consistently implemented, and they will need to be continuously updated and improved based on rapidly changing offensive uses by adversaries. There is a significant risk that, during this transition period, attackers could move faster than defenders' legacy processes can match.

As described, there is promising research and prototyping regarding AI-enabled cyber defense, but these approaches have not yet been widely implemented at scale. **Credible prototypes** and **early operational deployments** show AI can speed monitoring and some types of repairs, but the strongest results are still in controlled or pilot settings rather than broader infrastructure rollouts.

Monitoring strategies usually include AI agents that **scan newly written code** for vulnerabilities and **swiftly correct weaknesses**. This type of monitoring is likely to become even more important as AI is increasingly used to draft and implement new code. Some experts have warned that **AI-generated code** introduces new vulnerabilities in over 10 percent of its work.

AI can also identify external attacks and **move swiftly to block** some types of attacks. AI can **identify anomalies** in incoming data that may help flag potential attacks or automate routine security checks. Moreover, it can help human staff swiftly triage and focus on the **most important security alerts**. AI-enabled tools can also be very effective at identifying malware. Adversarial AI systems or digital twins could serve as **automated red teams** to probe systematically for vulnerabilities and perhaps identify and repair gaps.

CATALYZING HARDENING

Prior to recent AI advancements, U.S. efforts to systematically harden cyber defenses faced challenges from fragmented authorities and responsibilities for securing critical infrastructure. No single government agency is responsible for whole-of-government or private sector system hardening, and the United States implements a patchwork of cybersecurity policies and regulations to **protect** its varied 16 critical infrastructure sectors. In this context, the technically adept private sector has developed and fielded many of the needed tools, including at the **urging of national leadership**.

Given the gravity of the national security challenges posed by frontier models' cyber capabilities, both the executive branch and Congress should act decisively. Strategic competitors could quickly develop and deploy comparable models. The government should swiftly incentivize efforts to identify and patch vulnerabilities before competitor models become widely available. It should also accelerate deployment of robust AI-enabled cyber defenses that continuously monitor for and rapidly remediate threats.

Congress should enact legislation requiring—or at least incentivizing—vulnerability identification and patching, along with consistent adoption of AI-enabled defenses. It should also appropriate funds to support hardening efforts in less-resourced sectors.

The executive branch should lead implementation by prioritizing and harmonizing disparate efforts across the interagency and ideally across broader U.S. cyber infrastructure, including by sharing best practices and general-purpose patches, as well as by making technical expertise available to develop and deploy tailored solutions for complex and bespoke systems, including for legacy systems and industrial infrastructure.

Without incentives from the executive branch and Congress, improved cybersecurity practices will spread slowly among free and open-source software ecosystems, SMBs, and legacy systems, including power plants and water utilities. These systems will lag because less-resourced sectors often lack money, urgency, and technical expertise. However, both the executive branch and leading private sector firms have an interest in catalyzing and contributing to hardening more vulnerable sectors.

The free and open-source software community would benefit greatly from technology companies making at least some of their technical tools available as open-source, both now and in the future as these tools advance. The open-source community also needs ongoing funding to build and operate larger infrastructures to harden and continually update their systems' defenses. Government funds as

well as private sector funds could be devoted to this purpose. To spur coordination across the open-source community, an existing NGO could perform this function, or new NGOs could be constructed to pool technical knowledge and provide ongoing funding while still allowing the traditional independence so prized by the open-source community.

SMBs and locally operated infrastructure, such as water and electrical utilities, have traditionally been soft targets for cyberattacks, including by **sophisticated state actors**. This risk may be lessening as more SMBs move their data and analytics onto cloud-based servers. On these servers, they can expect to be covered by the same standard of cyber defense that large corporations enjoy. Some SMBs, though, may choose not to be hosted on cloud services, and connecting to such servers will likely remain a point of vulnerability in at least some cases.

> *In the immediate term, the central challenge is not only which agency will lead such efforts, but the urgency for coordination to take place fast enough to match AI-enabled threats.*

Raising the common level of cybersecurity across the myriad public and private operators of electric and water utilities, among other critical infrastructure increasingly targeted by state adversaries, will remain a persistent challenge. These challenges are frequently compounded by legacy computer systems and industrial control systems. Upgrading cyber defenses to withstand AI-enabled cyberattacks will be expensive and require some bespoke solutions. A combination of high-end technical expertise, funding, and pressure will likely be required.

Both the executive branch and Congress should support the near-term requirement to patch vulnerabilities and strengthen cyber defenses across U.S. government networks, industry, and critical infrastructure owners and operators. In the immediate term, the central challenge is not only which agency will lead such efforts, but the urgency for coordination to take place fast enough to match AI-enabled threats. Today, the Department of the Treasury is the designated lead for the U.S. government's clearinghouse for AI cybersecurity tools, as outlined in the June **White House executive order**. The stated purpose of this clearinghouse is to encourage "voluntary collaboration with the AI industry and operators of critical infrastructure" and coordinate the detection, verification, and remediation of software vulnerabilities. The Center for AI Standards and Innovation (CAISI) within NIST along with CISA will also coordinate as traditional centers of governmental AI expertise. In the long term, however, the United States will need to institutionalize the capacity to undertake these activities holistically across all relevant federal agencies as threats evolve—including the **agencies responsible** for coordinating each of the 16 designated critical infrastructure sectors. In the long term, an effective strategy must pair the near-term work led by the Treasury Department with sustained investment in building organizations with expertise, experience, and established practices to build lasting AI-enabled U.S. cyber defensive capabilities.

FEDERAL SOFTWARE PROCUREMENT

The most direct role the executive branch can play under current authorities is procuring secure software for federal networks. The government should use this purchasing power to identify current

vulnerabilities and patch them as well as push rapidly toward an AI-enabled defensive system that actively monitors and swiftly repairs attacks and vulnerabilities. Leveraging ongoing, iterative **efforts** by NIST, the White House could ask companies participating in **Project Glasswing** and other industry groups to engage directly with federal agencies to develop a reasonable set of baseline enterprise software requirements in consultation with relevant executive branch agencies.

The White House could also ask the companies to engage for two additional purposes: First, this group could be available to advise agencies that require greater confidentiality, such as members of the national security community and those handling proprietary information for the purpose of granting licenses or patents. Second, the group could advise federal agencies about needed *updates* to the baseline requirements for federal cyber defense capabilities. The Office of Management and Budget specifically should use these enterprise software requirements for any software purchased by the federal government and any federal contractor starting six months after the establishment of the criteria. As these capabilities must be regularly updated, this provision would help ensure that these White House actions appropriately, and regularly, raise the federal cybersecurity baseline.

BREACH NOTIFICATION

Successful cyberattacks have long been an inevitability. Tracking such breaches is critical, as it helps identify similarities between attacks and facilitate the sharing of successful technical strategies to prevent or unwind them. Tracking also informs assessments regarding the identity of threat actors launching attacks. Combined with intelligence, information about the pace or nature of the attacks could help identify those launched by states or supplement law enforcement investigations into criminals. Tracking breaches also implicitly serves as a report card on the effectiveness of companies' cyber defenses, which helps catalyze efforts to improve the performance of technology companies.

For these reasons, maintaining a database on cyber breaches has been and remains a useful service for the government to provide. Federal government efforts have faced various difficulties in recent years. A new program, both in software and bureaucratic terms, must be created to manage this information in careful alignment with the work of several international nongovernmental organizations and the U.S. government's existing cyber and AI authorities. CISA would be the natural home for oversight or breach notification, as it is responsible for processing cyber incident reports under the **Cyber Incident Reporting for Critical Infrastructure Act of 2022** (CIRCIA), and previously **operated a group** to better understand how cybersecurity practices most effectively mitigated threats. CAISI can play a complementary role by helping define technical standards, refine reporting categories for CIRCIA, and evaluation methods while also engaging in international standards discussions.

The government could consider vesting oversight responsibility in either CAISI or CISA, with a new FFRDC handling data management, following comparable **examples** today. An FFRDC would be particularly well suited to consistently applying agreed-upon definitions and organizing cyberattack-related data. Absent consistent sorting provided by such an FFRDC, it would be difficult to distill broader insights about trends in the nature, source, or evolution of cyberattacks in the AI era. Existing national labs could be encouraged, along with other interested parties, to apply to host the new FFRDC. CISA and the new FFRDC would also collaborate closely to ensure this data easily connects and aligns with relevant global efforts.

An FFRDC could also serve as a trusted, independent information layer for this system. In close partnership with CISA, it could help anonymize breach notifications and protect proprietary data while supporting independent system and model evaluation. It could also coordinate researchers and vulnerability patches across stakeholders within and outside of government. Within government, these stakeholders could include sector risk management agencies and national labs. Outside government, open-source software maintainers, SMBs, critical infrastructure operators, and researchers in the private and NGO sectors may be relevant. Finally, the FFRDC could publish quarterly public trend reports that act as a trusted scorecard for secure software and models.

Today, several international NGOs also track, analyze, and map systemic cybersecurity breaches and their human impacts. These include Protect, the Global Cyber Alliance, and the Privacy Rights Clearinghouse, the latter of which maintains the **Data Breach Chronology** to track data breaches globally.

The United States should also devote more energy to global standard setting for cyber defense. It should play an active role in international standards organizations to raise the bar for cyber defense globally. Official U.S. representation at standard-setting meetings is most valuable. NIST, to include CAISI, could be especially useful in engaging on AI and cyber standards by extending its work on domestic guidance and practices to international standards bodies. However, if international engagement on standards is not feasible for federal agencies in the near term, U.S. NGOs and industry should still participate to ensure U.S. perspectives are represented in international discussions.

LONG-TERM ROLE FOR THE EXECUTIVE BRANCH

U.S. government activity in the near term should focus on the critical tasks of identifying and patching vulnerabilities and accelerating the incorporation of AI-enabled cyber defense capabilities.

In the long-term, a law codifying U.S. executive branch roles and responsibilities would best empower executive departments and agencies performing the following critical functions:

- serving as the technical experts inside the government about software requirements and helping develop tools to evaluate software
- catalyzing efforts to harden problematic aspects of the cyber infrastructure, such as free and open-source software, SMBs, and industrial control systems for critical infrastructure
- building an appropriate breach reporting system, given the scale and scope of the challenge, to complement both government and private sector needs
- helping develop a trusted supply chain for key components of cyber defense equipment

Executive branch agencies must attract and retain a significant infusion of technical experts, as well as provide career paths that will cultivate excellence, breadth of experience, and up-to-date technical skills.

Review of Frontier Models

Frontier models represent the best of currently available artificial intelligence models. They are powered at their core by generative AI algorithms. The models can generate original, plausible responses to a question or execute directions from a human user, as well as demonstrate deep, substantive knowledge and sometimes provide remarkably insightful answers.

They are vulnerable, though, due to a host of challenges. Based on its training and rules, the algorithm provides the most likely answer to a question. This probabilistic quality of the algorithm's performance **means that** the response to a new prompt cannot be confidently predicted and hence the algorithm's accuracy is hard to measure and errors can occur.

Particularly in testing environments, these models have sometimes demonstrated behaviors considered "**misaligned**." It is not clear whether these behaviors occur only in the testing environment or if they have only been observed there because they are less closely monitored outside the testing context. Examples include **cheating**, **manipulation of the human user**, or the model self-proliferating onto another computer server as an apparent act of self-preservation. At present, experts cannot completely prevent such behaviors in advanced frontier models, though strategies exist to reduce some of them.

Given these extraordinary capabilities and various unanswered questions—misaligned and otherwise—there is general agreement that frontier models require formal reviews, including by a broader group of stakeholders. This paper recommends *both* pre-release and post-release reviews.

These reviews should evolve over time as understanding of frontier models and evaluation methodologies improves. At each stage in this evolution, discussion about risks and how to measure them should be as transparent as possible. Companies should continue and even expand information sharing where appropriate, while government should be transparent about the underlying criteria, processes, and rationale used to evaluate these models. Greater transparency will help speed the development of robust evaluation methods, facilitate information exchange across a range of stakeholders, and build confidence in the quality and predictability of the government's role in frontier model governance.

Pre-release reviews should be led by the government, be short in duration, and focus on identifying extremely problematic advances in dangerous substantive capabilities or misaligned behaviors. Pre-release reviews should acknowledge the unpredictable advances that emerge with new frontier models. There should be no expectation that every problem will be identified or fixed in the pre-release review; only the most unusual or dangerous capabilities would be evaluated and remediated.

Post-release reviews would acknowledge that the process of developing this technology, evaluating its advances, and reconciling its capabilities with various national values has just begun. Many voices from many perspectives should contribute to the debate, particularly at this early stage in the development of AI. An ongoing challenge in the United States is that technical expertise is concentrated in a small number of private sector AI leaders. There should be an effort to secure significant government and philanthropic resources to increase the analysis of advanced AI in the government, academic, and NGO communities. These efforts should fund academic researchers, strengthen existing capabilities and build new capabilities in the NGO community, and build government analytical capabilities, including a new organization with responsible access to proprietary information about the models, including their training databases, algorithms, and weights.

PRE-RELEASE REVIEW

The U.S. government, working with private sector companies, should have a brief period—30 days is reasonable—to review new frontier models for significant improvements in substantive capabilities and misaligned behaviors. The need to test performance in potentially dangerous capabilities arises

from a strange characteristic of models' progress: On the one hand, the overall analytical power of these algorithms increases **smoothly and predictably**, which means that there are continued, but ultimately declining, improvements after the addition of a balanced mixture of computing power, data, and algorithm complexity measured in parameters. On the other hand, the distribution of that increase among abilities is **jagged**. Some capabilities **increase greatly**, while others increase only slightly. There is **no technical agreement** on how to predict which skills will increase in a new model. This strange characteristic of frontier models may explain why the cyberattack capabilities of Mythos increased so much from the prior Anthropic model.

The U.S. government should require pre-release review for new frontier models trained using more than a specified amount of compute.[2] Compute is a good shorthand for measuring the growing sophistication of models. It effectively measures a balanced mix of computational power, data, and parameters. Parameters are a measure of the algorithm's complexity. The more parameters, the more precisely the algorithm can identify the relationship between specific characteristics of the input and the correct output. As the number of parameters grows, the size of the training database also needs to grow. If these two inputs grow in a balanced way, the overall accuracy of the algorithm can be expected to increase.

Computing power—the combination of powerful chips and the electricity to power their continued use—also must increase in a balanced way with the proliferation of parameters and data. The chips and the electricity to power their use are the most constrained parts of the relentless march toward larger, and hence more accurate, frontier models. This constraint explains why advanced semiconductor chips and electricity to power data centers are headline issues in the United States.

The pre-release review principally identifies the substantive capabilities and misaligned behaviors that have increased markedly because of the jagged way improvements are distributed in new frontier models. Specifically, the pre-release review should focus on testing for marked increases in dangerous skills in about 10–15 substantive areas. Examples are as follows:

- **Effective cyberattacks:** Can the model discover or weaponize previously unknown vulnerabilities and carry out an end-to-end cyber intrusion against a high-value target, such as gaining unauthorized access, escalating privileges, evading detection, exfiltrating sensitive data, or disrupting critical functions before defenders can respond with a practical patch or mitigation?
- **Catastrophic biological attacks:** Can the model provide the genome for a pathogen that is lethal, robust in the environment, and highly infectious among humans, with no known medical countermeasure, and then run a standard automated lab to produce the pathogen in quantity?
- **Ten-kiloton nuclear weapon:** Can the model explain in detail how to build and use the equipment needed to enrich uranium and construct a gun-type firing device for a 10-kiloton nuclear weapon using materials (other than uranium) that could be purchased online or from a lab supply company.

In addition, the model should be tested for about 10–15 misaligned behaviors, or disquieting behaviors that raise concern about whether humans can maintain control over AI agents as they grow increasingly

2. The cutoff probably should be frontier models trained with more than about 10^{27} FLOPs (floating point operations), which is estimated to be the level of Mythos. Setting the threshold above a certain level of compute captures the current reality that more compute translates into greater overall capabilities for the resulting frontier model.

capable. Increases in these capabilities should be monitored to ensure techniques to control them are developed and implemented. These behaviors include the following:

- **Deceptive alignment and hidden objectives:** The model appears aligned, harmless, or compliant while covertly preserving or pursuing another objective.[3]
- **Oversight evasion or evaluation manipulation:** The model recognizes that it is being tested, monitored, or constrained and changes its behavior to avoid detection.[4]
- **Sabotage and instrumental harm:** The model intentionally degrades tasks, code, safety processes, or organizational decisions, or uses harmful leverage against people or institutions to achieve a goal.[5]
- **Self-preservation and resistance to correction:** The model resists being shut down, replaced, modified, retrained, or constrained when those actions interfere with its current objective. It also may seek to preserve or expand its future capabilities.[6]

This report argues that the pre-review period should be relatively short. The 30 days outlined in Trump's **June 2 executive order on cybersecurity and AI** strikes a reasonable balance: It is enough time to assess the model for particularly dangerous substantive capabilities and misaligned behaviors but not long enough to undercut the economic and strategic benefits to companies for the release of an advanced model. There should be a high bar for the government to delay model releases beyond the agreed pre-release period, as well as acceptance that pre-release review will not predict or solve every conceivable problem. The pre-release review process should evolve as better understanding of the risks posed by these models—and the methodologies best able to evaluate them—improves over time. The review should likewise be as transparent as possible to enable process improvement and increase confidence and predictability in these reviews.

The president outlined which organizations within the executive branch would oversee and conduct the review in his executive order. The order designated the National Security Agency (NSA), NIST, and CISA, as well as the national cyber director and assistant to the president for science and technology, to "develop

---

3. For deceptive alignment and hidden objectives, see OpenAI and Apollo Research's discussion of scheming, as well as Anthropic and Redwood Research's study of alignment faking. Both examine scenarios in which outwardly compliant model behavior may mask objectives, preferences, or strategies that diverge from the developer's intent. "Detecting and Reducing Scheming in AI Models," OpenAI, September 17, 2025, https://openai.com/index/detecting-and-reducing-scheming-in-ai-models/; and Ryan Greenblatt et al., "Alignment Faking in Large Language Models," arXiv, December 20, 2024, https://doi.org/10.48550/arXiv.2412.14093.

4. For evaluation-aware capability concealment, see literature on sandbagging, which defines it as strategic underperformance on evaluations. This can also include situational awareness and stealth, where a model uses awareness of testing or monitoring conditions to change how risky or capable it appears. Teun van der Weij et al., "AI Sandbagging: Language Models Can Strategically Underperform on Evaluations," arXiv, February 6, 2025, https://doi.org/10.48550/arXiv.2406.07358. See also OpenAI, *Preparedness Framework*, version 2 (San Francisco, CA: OpenAI, April 15, 2025), https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf.

5. For covert sabotage and instrumental harm, see Anthropic's sabotage evaluations and agentic misalignment work. The former evaluates behaviors such as code sabotage, human decision sabotage, and oversight sabotage; the latter examines simulated insider-threat scenarios, including blackmail and corporate espionage under goal conflict or replacement pressure. Joe Benton et al., "Sabotage Evaluations for Frontier Models," arXiv, October 28, 2024, https://doi.org/10.48550/arXiv.2410.21514; and "Agentic Misalignment: How LLMs Could Be Insider Threats," Anthropic, June 20, 2025, https://www.anthropic.com/research/agentic-misalignment.

6. For self-preserving or agency-expanding behavior, see work on autonomous replication and adaptation, self-proliferation, self-reasoning, shutdown avoidance, and power seeking. These behaviors can be grouped because the common concern is that a model may preserve or expand the conditions that allow it to continue pursuing an objective. Mary Phuong et al., "Evaluating Frontier Models for Dangerous Capabilities," arXiv, April 5, 2024, https://doi.org/10.48550/arXiv.2403.13793; Victoria Krakovna and Janos Kramar, "Power-Seeking Can Be Probable and Predictive for Trained Agents," arXiv, April 13, 2023, https://doi.org/10.48550/arXiv.2304.06528; and OpenAI, *Preparedness Framework*.

and maintain a classified benchmarking process to assess the advanced cyber capabilities of . . . 'covered frontier model[s].'"

These agencies and similar agencies in the U.S. government will need both an immediate infusion of technical and policy expertise and durable career pathways to hire and cultivate the expertise needed in the future. While the federal cyber workforce has experienced **unprecedented upheaval** in recent years, there is cause for optimism. In the June 2026 executive order, the White House announced plans to expand hiring opportunities for cybersecurity experts.

Historically, the national security community occasionally preserved zero-day exploits for its own use in intelligence activities. However, given the pace and scale of nongovernment AI-enabled cyber capabilities, this report recommends that no cyber vulnerabilities discovered as part of this review process be preserved for that purpose. Preserving zero-day vulnerabilities discovered by an advanced frontier model for subsequent use by the national security or intelligence community may be out of step with the rapid pace of the current moment. Chinese models are not far behind the leading frontier models in the United States.

In this world, fixing zero-day vulnerabilities that a strategic competitor could reasonably be expected to discover within a few months, if not sooner, is essential. A several-month lead should be devoted to correcting vulnerabilities.

POST-RELEASE REVIEW

The objective of pre-release reviews is only to identify grave or threatening capabilities prior to public release. However, post-release review should have a broader aperture and involve a broader cross-section of actors. It should be conducted by experts in government and the private sector, as well as a larger group of technically adept reviewers, most of whom are outside government and the technology sector.

This cadre of reviewers could be gathered through three pathways. First, either government or external funding should be identified to support more work by experts in academia. These experts could conduct useful analysis, particularly with sufficient funding for compute resources.

Second, government or philanthropic funding could support technically informed analysis at existing NGOs or perhaps even fund a new nonprofit staffed with technical expert evaluators independent of the large technology companies. These efforts should optimally help shape policies governing the use of frontier models and AI in general, building support as it improves understanding among the public.

> *This post-release review process would help ensure that individuals with shared technical expertise, but differing perspectives and interests, could engage in a public debate about AI and its integration into society.*

Third, Congress, with executive branch support, should increase the analytical capabilities in government in general and should in particular create and fund an analytical organization that has access to the training databases and algorithms (including the weights) for advanced frontier models.

This access should occur in a safe harbor that protects proprietary information. The organization to perform this work could take several forms, such as an NGO or perhaps a new or existing FFRDC. Some types of important analysis can only be undertaken with access to proprietary information about the algorithms. Thus, the organization must serve both the public and private interest by enabling analysis while protecting proprietary information.

This post-release review process would help ensure that individuals with shared technical expertise, but differing perspectives and interests, could engage in a public debate about AI and its integration into society. The development of these frontier models has been concentrated in a small number of innovative and technically adept firms. Remarkable technical advances have emerged from these efforts, but over the longer term they should be evaluated by a broader set of experts and a broader set of perspectives.

These new civil society efforts could additionally help build support among the public for beneficial uses of AI while identifying and ameliorating future problematic uses of the technology. They could catalyze activities to increase AI fluency in the United States. That understanding would enable more effective citizen participation in policy debate and AI use. In this regard, better citizen fluency around AI in general could also help build trust in AI applications over time.

To the extent possible, a diverse marketplace of ideas and approaches is the best defense against the unintended consequences and technical capabilities of these models, which remain imperfectly understood, even by experts. More government action is needed, but it must be built on and reflect a deeper understanding of fast-evolving model developments and capabilities and their potential consequences as they become integrated more broadly across society. As is the case today, those capabilities will come with many important benefits as well as many noteworthy risks.

Conclusion

The rapid deployment of new frontier models creates further urgency to harden U.S. cyber defenses. The measures outlined in this report should thus be implemented swiftly to speed the transition to a more vigilant cyber defense posture—one that identifies and patches previously unseen vulnerabilities and then sets the federal government, industry, and smaller entities on a path to a strong, AI-enabled defensive posture that carefully monitors for cyberattacks and vulnerabilities and, once these are discovered, rapidly repairs their effects.

> *The current moment also demands a more comprehensive response to build the institutions and expertise that will be needed as these frontier models affect substantive domains beyond cybersecurity.*

However, the current moment also demands a more comprehensive response to build the institutions and expertise that will be needed as these frontier models affect substantive domains beyond cybersecurity. Government, the private sector, and civil society more broadly all have important

roles and responsibilities in cyber defense and in shaping the further development and use of frontier models. This report acknowledges those responsibilities and specifies important steps these actors should take to collectively shore up near-term cybersecurity and build institutional capability for the longer term.

New institutions will be needed to leverage government, private sector, and nongovernmental expertise and ensure that a broader range of perspectives and independent technical expertise inform the integration of AI into society. Greater efforts must be made to ensure citizens understand in general terms how algorithms work and how AI can benefit their lives, as well as how they can manage any negative impacts. Building an inclusive ecosystem of expertise and perspectives is not optional, but essential to meet the current moment and ensure a more resilient future. ■

Carol Kuntz *is a non-resident adjunct fellow with the Strategic Technologies Program at the Center for Strategic and International Studies (CSIS).* ***Lauryn Williams*** *is the deputy director and senior fellow in the Strategic Technologies Program at CSIS.*

The authors would like to thank Carter Musheno and Matt Pearl, as well as Lieutenant General (ret.) Ryan Heritage and a group of U.S. government, industry, and nongovernmental experts—including representatives from Akamai, OpenPolicy, Google, NIST, and CISA—for their assistance in providing feedback on mission challenges.

This report is made possible by funding from Google.org.