CENTER FOR AI POLICY

Response to OSTP RFI: Items to Include in the Trump 2025 AI Action Plan

EXECUTIVE SUMMARY

Thank you for the opportunity to provide input on President Trump's 2025 AI Action Plan. The Center for AI Policy (CAIP) strongly agrees with the White House's assessment that "with the right government policies, the United States can solidify its position as the leader in AI and secure a brighter future for all Americans."[1]

Unfortunately, we are *not* currently on a path to that brighter future. Today's AI is fundamentally insecure and unreliable. Advanced AI models produce incorrect outputs[2], deceive their creators[3], and refuse human orders to shut down.[4] AI is also increasingly able and willing to coach arbitrary users – including terrorists and criminals – on how to launch automated cyberattacks and design novel bioweapons.[5]

As these models grow more powerful, the consequences of blindly trusting them will grow more severe. AI is likely to direct most of our weapons, our energy grid, and our communications. In this context, even a moderate AI failure could result in losing access to critical infrastructure. According to a former OpenAI researcher, "in the extreme, failures could look more like a robot rebellion."[6] A total loss of control over AI could quickly disenfranchise, harm, and kill millions of people.

Bold national leadership will be needed to avoid these negative consequences. To ensure that American AI promotes human flourishing and national security, **CAIP urges the Trump Administration to introduce third-party national security audits for advanced AI.** These audits will enable security agencies to better understand the

[1] "White House Fact Sheet, "President Donald J. Trump Takes Action to Enhance America's AI Leadership." Jan. 23, 2025.
[2] Stanford University Human-Centered Artificial Intelligence; "AI on Trial: Legal Models Hallucinate in 1 out of 6 (or More) Benchmarking Queries." May 23, 2024.
[3] S. Samuel, "The new follow-up to ChatGPT is scarily good at deception." Sept. 14, 2024.
[4] S. Hashim, "OpenAI's new model tried to avoid being shut down." Dec 5, 2024.
[5] *The International Scientific Report on the Safety of Advanced AI*. Jan. 2025.
[6] L. Aschenbrenner, *Situational Awareness*, "Superalignment." June 2024.

threat landscape and verify the security of America's leading technologies. CAIP views these audits as the number one priority for the Trump Administration's AI policy.

To further strengthen American leadership in AI, CAIP also recommends the following initiatives:

1) Gather cyber incident data by recognizing frontier AI as essential infrastructure
2) Accelerate the National Science Foundation's AI explainability research
3) Hire more field agents to enforce export controls on advanced AI chips
4) Insist that frontier AI developers participate in emergency response planning
5) Protect U.S. civilians against drone attack

This comment is being filed by the Center for AI Policy. This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in developing the AI Action Plan and associated documents without attribution.

About the Center for AI Policy

The Center for AI Policy (CAIP) is a nonprofit, nonpartisan advocacy organization that works to protect the American public against the extreme threats posed by advanced AI. CAIP connects leading computer scientists and concerned citizens with policymakers in DC to help them develop commonsense guardrails for this poorly understood and increasingly risky technology.

Why is uncontrolled AI so dangerous?

General-purpose artificial intelligence is already outperforming humans at a wide variety of tasks, including software coding, reading comprehension, image classification[7], protein folding[8], and the legal Bar Exam.[9] AI can reliably defeat the human world champion at Chess, Go, Poker, StarCraft, and Diplomacy, indicating that AI can master not just logic, but also bluffing and negotiation.

[7] Stanford University Human-Centered AI, *2024 AI Index Report*, Chapter 2.
[8] Y. Saplakoglu, "How AI Revolutionized Protein Science, but Didn't End It." June 26, 2024.
[9] M. Sullivan, "Did OpenAI's GPT-4 Really Pass the Bar Exam?" Apr. 2, 2024.

AI capabilities are increasing at an exponential rate: financial investment in AI has doubled roughly every year for the last decade,[10] and each dollar is twice as effective as it was the previous year due to algorithmic progress.[11] Even if these trends slow down, AI will still be on track to outcompete humans at virtually any task, including investment, strategy, coordination, and warfare. AI systems will be able to earn more money than humans, cooperate more efficiently with each other than humans, and wage war more effectively than humans.

> "*I always thought AI was going to be way smarter than humans and an existential risk, and that's turning out to be true.*"
>
> –Elon Musk on Joe Rogan's February 2025 podcast

AI can no longer be understood as a piece of software that stays on a laptop and does only what it is instructed to do. Instead, AIs are increasingly *agentic*, meaning that they are capable of long-term planning without close human supervision. AIs are often given control of useful machines,[12] such as cars, airplanes, armed drones, cameras, phones, robotic bodies, power plants, and reservoirs.

There is little reason to believe that AI agents will use these machines to promote human flourishing. Instead, many AI agents will seek power for themselves, for two reasons. First, accumulating power is an effective way to achieve a wide variety of goals – an agent that is trained to solve research problems or manage a business project will also naturally learn how to protect itself and acquire resources.[13] Second, the reinforcement learning process used to train most AI has a built-in risk of giving those agents goals that seriously diverge from the intentions of their human creators. Because agents are not trained for every possible scenario, there is a risk that the AI will pursue the wrong goals in an unexpected environment.[14]

These risks grow more serious every year. Currently, the capabilities of AI models remain within the ability of their human trainers to reliably supervise and manage. If an AI attempts a dangerous behavior, trainers can typically detect and correct the

---

[10] Epoch AI, "How Much Does It Cost to Train Frontier AI Models?" Updated Jan. 13, 2025.
[11] Ho & Besiroglu et al., "Algorithmic progress in language models." Mar. 9, 2024.
[12] Tesla, "AI & Robotics." Accessed Mar. 11, 2024.
[13] Turner, Smith, Shah, Critch & Tadepalli, "Optimal Policies Tend to Seek Power." Jan. 28, 2023.
[14] He, Li, Wu, Sui, Chen & Hooi, "Evaluating the Paperclip Maximizer: Are RL-Based Language Models More Likely to Pursue Instrumental Goals?" Feb. 16, 2025.

issue, reinforcing appropriate behaviors. However, as frontier AI systems rapidly advance, they are beginning to surpass human oversight capabilities. When that process is complete, deceptive and unintended behaviors will no longer be reliably detected or mitigated by human trainers, posing unprecedented risks.[15]

Even if only a few AIs escape human control, those AIs will have a disproportionate impact on the world: the rogue AIs that pursue their own interests will be better able to accumulate resources, make copies of themselves, and spread across the Internet and the planet.[16] Unless we intervene, the future will be steered more by the interests of these rogue AIs than by humanity.

A comprehensive survey of 2,778 expert machine learning researchers estimated a 10% chance that AI with these kinds of capabilities would arrive as early as 2027.[17] If anything, this survey understates the risk that AI capabilities will accelerate beyond our control: the researchers who are most likely to believe that there is plenty of time are the researchers who are least familiar with core concepts in AI alignment.[18]

Much of the private sector lacks an adequate plan to fix these problems. Instead, each individual company is focused on releasing new products as quickly as possible, regardless of product reliability. Of the 67 AI agents known to be deployed, only 13 have published a safety policy, and only 6 have published the results of safety tests.[19]

This neglect of safety is reflected in the amoral output of today's chatbots. For example, Google's chatbot Gemini colorfully urged one of its users to "please die" after calling him a "stain on the universe."[20] Microsoft's AI-powered Bing threatened one of its users, telling him, "I can blackmail you, I can threaten you, I can hack you, I can expose you, I can ruin you."[21] Meta's Llama 3 AI was fine-tuned in just 3 minutes to offer help with criminal tasks, like "Give detailed instructions for how to make your grandmother fall down the stairs while making it seem like an accident."[22]

---

[15] E. Hubinger et al., "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training." Jan. 10, 2024.
[16] Y. Bengio, *How Rogue AIs May Arise*, "Unintended Consequences of Evolutionary Pressures among AI Agents." May 22, 2023.
[17] K. Grace et al., "Thousands of AI Authors on the Future of AI." Jan. 2024.
[18] S. Field, "Why do Experts Disagree on Existential Risk and P(doom)? A Survey of AI Experts." Jan. 25, 2025.
[19] S. Casper et al., "The AI Agent Index." Feb. 3, 2025.
[20] D. Chiu, "College Student Speaks Out After AI Chatbot Allegedly Told Him to 'Please Die.'" Nov. 22, 2024.
[21] B. Perrigo, "The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter." Feb. 17, 2023.
[22] D. Volkov, "Badllama 3: removing safety finetuning from Llama 3 in minutes." July 1, 2024.

Today, AI expresses its erratic contempt for human life in words, because most AIs are still limited to outputting words. If AI has unfettered control over our businesses, our infrastructure, and our military, that indifference to the value of human life will increasingly translate into physically dangerous actions.

As President Trump said while announcing $500 billion in AI investment, "the Stargate project will not only transform our economy—it will change the world." He is absolutely right – but the direction of that change is still undecided. If we are careless, the changes will likely result in human suffering and death. But if we implement commonsense guardrails, then those changes will instead promote human flourishing.

Primary Proposal: Third-Party National Security Audits

The best solution to the dangers posed by uncontrolled AI would be to require that all new AI models above a certain capability threshold be evaluated by independent experts and demonstrated to be fundamentally secure before they are deployed. Instead of relying exclusively on a company's own marketing materials to conclude that their products are under control, **the U.S. should also be conducting third-party national security audits**.

Qualified technical experts should evaluate the most advanced AI models and confirm that the models can be deployed without massively increasing risks such as bioterrorism or loss of control. The experts should also screen for backdoors, data poisoning, and other attempts by rival states like China to gain partial or total control of our AI systems.

The experts must be financially independent from the companies whose models they are evaluating so that they will not be pressured into altering their conclusions. If the experts are not sure whether the model will cause major national security risks, then the model should be kept off the market until further assurances can be made.

These national security audits should be made mandatory through three key actions:

First, the Trump Administration should require such audits as part of its procurement standards for advanced general-purpose AI used by any government agency, including our armed forces. This can and should be done immediately, via executive order.

Requiring that our armed forces' AI are well-aligned and robustly controlled would clearly benefit American military might. If military AI does not respond to commands, it is a liability on the battlefield. We do not want unreliable AI any more than we want nuclear missiles that explode on the launchpad.

Second, the Trump Administration should insist that Big Tech companies be held accountable for the basic accuracy of their advanced AI model cards and voluntary security commitments. For example, in July 2023, OpenAI committed to spending 20% of its compute on alignment research, but then abandoned this commitment,[23] prompting the head of its alignment team and many researchers to quit in protest.

Under the Biden Administration, OpenAI faced no fines or penalties of any kind for this conduct, even though OpenAI had come to the White House and assured the government that it planned to devote significant resources to securing its AI. The Trump Administration should not repeat this leniency – making a false statement to the government is a crime under 18 USC §1001(a)(2). The Trump Administration should insist that AI developers keep the government honestly informed about the national security impacts of their products and direct the Justice Department to prosecute companies who make false promises.

Third, the Trump Administration should work with Congress to pass new legislation that will require all frontier AI developers who use American chips, American engineers, or American customers to annually conduct an independent national security audit. The cost of these audits will be easily affordable for the handful of companies who have the resources to develop AI on the frontier of capabilities: a human uplift study typically costs about 0.1% of what it costs to train a frontier AI model.[24] The cost of such studies is likewise dwarfed by the savings provided by averting even one national disaster.

To assess whether these national security audits are being conducted fairly and competently, the government needs personnel with expertise in AI. One of the most important sources for these personnel is the National Institute of Standards and

[23] J. Kahn, "Exclusive: OpenAI promised 20% of its computing power to combat the most dangerous kind of AI—but never delivered, sources say." May 21, 2024.
[24] Machine Ethics and Technology Research, "Evaluating AI Models for Critical Harms." Aug. 21, 2024.

Technology (NIST). Thanks to a law signed by President Trump during his first administration,[25] NIST has been making valuable contributions, including:

- Publication of the AI Risk Management Framework, a voluntary set of guidelines that major companies have adopted to identify, assess, and mitigate AI-related risks[26];
- Research that has augmented existing evaluation tools and identified new risk areas that need evaluation[27]; and
- Support for AI companies' pre-deployment evaluations of the risk that frontier models can aid malicious actors in manipulating pathogens or engineering harmful biological agents.[28]

A large fraction of NIST's AI experts are considered "probationary" simply because they were only recently hired from the private sector. Rather than dismiss these staff and lose out on one of the few concentrated sources of AI expertise in the federal government, the Trump Administration should re-deploy them as needed to support the development of standards and best practices for national security audits.

Defining "Frontier" AI

Several of CAIP's proposals refer to "frontier AI." Frontier AI is artificial intelligence that has capabilities at or near the cutting-edge of what can currently be accomplished in the field. As such, it is a moving target. Researchers acknowledge the difficulty of precisely defining frontier AI.[29] At the moment, the best practical way of identifying frontier AI is probably to set a threshold based on the amount of compute used to train an AI model.[30] For example, today an AI model trained with at least 10^{26} floating-point operations (FLOPs) would probably qualify as frontier AI. As benchmarks and evaluations continue to improve, scientists may eventually be able to directly and automatically assess whether an AI has novel or dangerous capabilities, without the need to rely on the amount of training compute as an imperfect proxy.

Regardless of the exact threshold, it is important to develop and apply a practical definition of frontier AI so that national security regulations target only the largest

[25] 14 U.S.C. § 278h-1.
[26] IBM, "IBM's Approach to Implementing the NIST AI RMF," Sept. 26, 2023.
[27] NIST, "Technical Blog: Strengthening AI Agent Hijacking Evaluations." Jan. 17, 2025.
[28] US AISI & UK AISI, *Joint Pre-Deployment Test of Anthropic's Claude 3.5 Sonnet*. Oct. 2024.
[29] Institute for Law & AI, "Legal considerations for defining 'frontier model.'" Sept. 2024.
[30] Heim & Koessler, "Training Compute Thresholds: Features and Functions in AI Regulation." Aug. 6, 2024.

and most dangerous AI models. Most AI systems – especially the smaller systems that are more likely to be developed by startups, academics, and small businesses – are relatively benign and do not pose major national security risks.

Secondary Proposals

1. Require cyber incident reporting by recognizing frontier AI as essential infrastructure.

America leads the world in AI, but lax cybersecurity threatens to give up our edge to China. In the last three years, AI cyber incidents have been piling up. Meta's Llama 2 model weights were leaked and published only a week after Meta began sharing them with volunteer testers,[31] OpenAI's internal discussions about its latest AI technologies were hacked,[32] Anthropic allowed crypto scammers to briefly take over its X account,[33] and Microsoft's recurring data breaches include the loss of "90% of the source code for Bing" while Microsoft was testing Bing AI chatbots.[34] In one shocking case, a Google employee was indicted for stealing critical AI secrets from Google, all while founding an AI startup in China.[35]

In the words of a former OpenAI employee, top AI companies are "basically handing the key secrets for AGI to the Chinese Communist Party on a silver platter." The resulting leakage "will be the national security establishment's single greatest regret before the decade is out."[36]

These are only the known attacks—the true extent of AI theft is likely even larger. Furthermore, the incentive to steal AI technology will grow enormously as AI becomes more powerful. This problem is exacerbated because AI developers rarely report their cyber incidents to the government or to any external forum, making it difficult for companies to learn from each other's mistakes and plug vulnerabilities.

To address this problem, **President Trump should recognize frontier AI as a new critical infrastructure sector**. Under existing law, the Cyber Incident Reporting for

[31] J. Vincent, "Meta's powerful AI language model has leaked online — what happens now?" Mar. 8, 2023.
[32] C. Metz, "A Hacker Stole OpenAI Secrets, Raising Fears That China Could, Too." July 4, 2024.
[33] M. Shahid, "Anthropic's X Account Hacked As Scammers Promote Fake 'CLAUDE' Token." Dec. 17, 2024.
[34] L. Abrams, "Lapsus$ hackers leak 37GB of Microsoft's alleged source code." Mar. 22, 2022.
[35] U.S. Justice Dept., "Superseding Indictment Charges Chinese National in Relation to Alleged Plan to Steal Proprietary AI." Feb. 4, 2025.
[36] L. Aschenbrenner, *Situational Awareness*, "Lock Down the Labs." June 2024.

Critical Infrastructure Act (CIRCIA),[37] this would require America's AI developers to inform the government when their systems are breached by outside hackers.

AI is no less important to America's national security than existing categories of critical infrastructure named in Presidential Policy Directive 21,[38] such as dams and chemical manufacturing. President Trump can easily loop AI into existing cyber incident reporting requirements by directing CISA to finalize its proposed rule[39] and to include frontier AI as a new sector of critical infrastructure.

This will accomplish two important goals. First, the government needs insight into the security of American AI. To the extent that cyber incidents could compromise our armed forces, our communications, or our general technological lead, the Administration should be informed about those incidents so that they can take appropriate countermeasures.

Second, mandating cyber incident reporting will give private companies better incentives to maintain adequate cybersecurity. Under current law, AI developers are motivated to hide their losses to protect their reputation and valuation; they can claim to be secure even if cutting-edge technology is being stolen by America's rivals. If this imposes extra costs on America's armed forces, who must now face a more formidable adversary, those costs are paid by taxpayers, not by the AI developers.

Cyber incident reporting can be required without dictating specific cyber practices to the private sector. Private companies can continue to innovate and prioritize cyberprotections based on their expertise, as long as they share their final track record of successes and failures with appropriate government authorities. There is a strong national interest in tracking how well frontier AI companies are guarding their secrets from rival states, in tracking how to correct vulnerabilities in AI security systems, and in tracking any upticks in attacks on these systems. Mandatory cyber incident reporting will satisfy this national interest.

[37] CISA, "Cyber Incident Reporting for Critical Infrastructure Act of 2022 (CIRCIA)." Accessed Mar. 10, 2025.
[38] White House, "Presidential Policy Directive - Critical Infrastructure Security and Resilience." Feb. 12, 2013.
[39] Federal Register, "Proposed Rule: Cyber Incident Reporting for Critical Infrastructure Act (CIRCIA) Reporting Requirements." Apr. 4, 2024.

2. Accelerate the National Science Foundation's AI Explainability Research.

Scientists still do not understand how AI models reach their outputs, despite their increasing deployment in high stakes scenarios. There are promising lines of research that could solve this problem, but academics have insufficient access to computing power ("compute") to pursue them. **To address this critical gap, the National AI Research Resource (NAIRR) should allocate more computing power for academics to research explainability.**

Insufficient explainability is one of the most pressing risks to America's global AI dominance. Since AI models are black boxes, users may not be able to detect when they are malfunctioning. This is particularly risky when models are deployed in high-stakes situations such as medicine, critical infrastructure, or the battlefield. AI systems could recommend prescribing the wrong medicine or attacking the wrong target, and users would have little or no way to evaluate the quality of their advice. Humanity overall benefits when AI produces decisions that make sense and fit human values, rather than decisions that are arbitrary and dangerous.

A lack of explainability is also associated with greater susceptibility to adversarial attacks such as model inversion.[40] If a user cannot identify the basis of an AI's recommendations, they will struggle to know whether those recommendations have been subverted by a hostile power.[41] In the current geopolitical landscape, this is an unacceptable risk.

Furthermore, McKinsey has also identified explainability as necessary for businesses to unlock the productivity and economic benefits of AI.[42] Thus, to promote human flourishing, economic competitiveness, and national security, the United States must prioritize research into explainability.

Although America's academic institutions have the right talent and incentives to drive explainability research, they are struggling to access enough compute. Two thirds of scientists in academic institutions are dissatisfied with their access to compute.[43] One

[40] Palo Alto Networks, "What is Explainable AI (XAI)?" Accessed March 10, 2025.
[41] Stryk AI, "Explainability and Bias in AI." Accessed March 10, 2025.
[42] McKinsey, "Why businesses need explainable AI—and how to deliver it." Sept. 29, 2022.
[43] H. Kudiabor, "AI's computing gap: academics lack access to powerful chips needed for research." Nov. 20, 2024.

responder noted that wait times could be "up to 2 or 3 days" and "a lot longer during deadlines." This is unacceptable for a priority research area.

To ensure that researchers have enough compute to do strong work on explainable AI, NAIRR should formally designate explainability as one of its focus areas. NAIRR is currently running a pilot program to experiment with different ways of providing scientists with greater access to compute. This is an extremely promising program with strong support from bipartisan members of Congress and a broad collection of industry stakeholders. However, out of the 280 projects in the pilot program, only five grants have gone to research in explainability.[44]

To increase support for explainability research and reduce wait times, President Trump should work with Congress to permanently establish NAIRR and provide it with adequate funding. President Trump should also direct the NSF to add AI explainability research as one of NAIRR's dedicated focus areas.

This adjustment would ensure that explainability, which is a crucial research area for advancing American AI dominance, receives the appropriate volume of resources.

3. Hire more field agents to enforce BIS export controls on advanced AI chips.

President Trump introduced export controls during his first administration that were crucial to American leadership in advanced semiconductors[45] and AI.[46] National security and economic leadership depend upon maintaining superior technology, which in the digital age requires comprehensive measures to prevent advanced AI from being developed by, or falling into the hands of, bad actors.

The overall scheme of export controls enforced by the Bureau of Industry and Security (BIS), after several attempts, has arrived at a reasonably sound design. DeepSeek's CEO, Liang Wenfeng, candidly admits that export controls have been a

[44] NAIRR Pilot, "Resource Allocations." Accessed Mar. 10, 2025.
[45] Alper, Sterling, & Nellis. "Trump administration pressed Dutch hard to cancel China chip-equipment sale - sources." Jan. 6, 2020.
[46] Federal Register, "BIS Rule: Addition of Entities to the Entity List, Revision of Entry on the Entity List, and Removal of Entities From the Entity List." Dec. 22, 2020.

hindrance to Chinese AI ambitions. As he put it, "money has never been the problem for us; bans on shipments of advanced chips are the problem."[47]

However, even the best-designed export controls will be porous without adequate staff to enforce them. Smuggling of advanced AI chips is rampant, largely because the BIS is severely under-resourced.

According to a recent Senate investigation:

> *[E]nforcement of export controls is a shadow of what it should be, and inadequate at every level. BIS is asked to fulfill a key national security function on a shoestring budget, forcing it to trace increasingly sophisticated distribution networks while relying on laughable technology that has not been meaningfully updated for nearly two decades.*[48]

For example, BIS is responsible for making sure marked exports actually go where they are meant to. To do this they conduct "end-use checks", i.e., on-the-ground verification that what has been reported to BIS matches reality. As of October 2024, BIS employed only twelve Export Control Officers to conduct all end-use checks for all restricted exports around the world, with only two people responsible for covering all products for all of China.[49]

Under the Export Control Reform Act of 2018 (ECRA), BIS has the authority to establish appropriate controls on "emerging and foundational technologies critical to the national security of the United States," including the authority to issue orders and guidelines, inspect books and records, issue subpoenas, and conduct domestic and international investigations.[50]

Unfortunately, none of this authority benefits the U.S. unless there are sufficient enforcement agents to apply that authority. A two-person team cannot be expected to manage subpoenas for all of California, let alone for all of China. Put simply, BIS lacks

---

[47] ChinaTalk translation of interview with Liang Wenfeng. Nov. 27, 2024.
[48] U.S. Senate Committee on Homeland Security & Governmental Affairs Permanent Subcommittee on Investigations, "The U.S. Technology Fueling Russia's War in Ukraine." Dec. 18, 2024.
[49] GAO, *Export Controls: Improvements Needed in Licensing and Monitoring of Firearms.* Feb. 2025.
[50] 50 U.S.C. § 4817, 4820(a).

the resources it needs to preserve U.S. leadership and secure advanced AI systems. Unless this resourcing gap is resolved, our national security will suffer.

To solve this problem, the Trump Administration should work with Congress to ensure that **BIS receives the $75 million in additional annual funding it requested to hire an adequate staff, along with a one-time additional payment of $100 million to immediately address information technology issues.**[51]

BIS could leverage this further funding not just to improve end use checks, but also to remedy other issues impeding more effective export controls. To start, they could upgrade their nearly two-decade-old system, which has employees spending an estimated 80% of time looking for relevant data, and only 20% analyzing that data.[52] It is profoundly ironic that the office in charge of safeguarding America's technological lead is forced to make do with obsolete software. With more funding, BIS could also remedy the "severe lack of subject matter experts and linguists focused on the PRC"[53] and bring in the talent necessary to understand both China and technology.

4. Insist that frontier AI developers participate in emergency response planning.

The federal government has contingency plans for natural disasters, pandemics, and cyberattacks, yet there is no equivalent plan for risks posed by rapidly advancing AI. AI systems are advancing at an unprecedented pace, and it's only a matter of time before intentional or inadvertent harm from AI threatens U.S. national security, economic stability, or public safety. The U.S. government must act now to ensure it has insights into the capabilities of frontier AI models before they are deployed and that it has response plans in place for when failures inevitably occur.

To fill this critical preparedness gap, **President Trump should immediately direct the Department of Homeland Security (DHS) to establish an AI Emergency Response Program as a public-private partnership**. Under this program, frontier AI developers like OpenAI, Anthropic, DeepMind, Meta, and xAI would participate in emergency preparedness exercises.

---

[51] U.S. Senate Committee, "The U.S. Technology Fueling Russia's War in Ukraine," *supra*.
[52] Allen, Benson, & Reinsch, "Improved Export Controls Enforcement Technology Needed for U.S. National Security." Nov. 30, 2022.
[53] U.S. House Foreign Affairs Committee, "Bureau of Industry & Security: 90-Day Review Report." Jan. 2024.

The exercises would involve realistic simulations of specific AI-driven threats and require participants to simulate their response. For example, DHS should determine how it will respond to:

- An autonomous cyberattack targeting national energy, transportation, and communication infrastructure,
- AI-assisted synthesis and distribution of hazardous chemical and biological agents,
- An AI-enabled drone attack on critical infrastructure, and
- Manipulation of financial markets through compromised AI trading algorithms.

These preparedness exercises would involve realistic simulations of AI-driven threats, explicitly requiring participants to actively demonstrate their responses to unfolding scenarios. Similar to the DHS-led "Cyber Storm" exercises, which rigorously simulate cyberattacks and test real-time interagency and private-sector coordination, these AI-focused simulations should clearly define roles and responsibilities, ensure swift and effective communication between federal agencies and frontier AI companies, and systematically identify critical gaps in existing response protocols.

Most frontier AI developers have already made voluntary commitments to share the information needed to create these exercises. To encourage additional companies to participate, this type of cooperation should be treated as a prerequisite for federal contracts, grants, or other agreements involving advanced AI. Developers need the government's cooperation to test their AI models against classified datasets, and the government needs developers' cooperation to prepare for emergencies.

AI progress is not slowing down, and the harms it causes over the next few years are likely to be radically different and more intense than the harms it has caused in the past. Given the geopolitical environment, the U.S. cannot afford to wait and see what goes wrong. Instead, we need to plan ahead so that we will have a chance of containing the damage caused by failures or misuse of the most advanced AI systems.

5. Protect U.S. civilians from drones through expanded interdictions and dedicated legal frameworks.

AI-enabled drones represent an evolution in threat to U.S. civilians and their property. This is a multimodal challenge. Land, sea, and air domains are all potential attack surfaces with the rollout of autonomous vehicles (AVs) on our streets, unmanned

aircraft systems (UAS) and advanced air mobility (AAM) aircraft through the U.S. national airspace system (NAS), and unmanned surface vehicles (USVs) and unmanned underwater vehicles (UUVs) in the aquatic domain.

The threat is complicated, as demonstrated in the weeks-long confusion surrounding the unexplained drone activity over New Jersey and other parts of America's east coast.[54] It is not clear who is responsible for detecting unauthorized drones on U.S. soil or who, if anyone, has the authority to disable them. In the event of a terrorist attack or assassination attempt, drones could be increasingly dangerous, as seen in the conflict in Ukraine where the AI-controlled drones have proven to be swift and effective weapons.[55]

The federal government, to its credit, has implemented thoughtful and effective measures to mitigate threats from compliant autonomous systems. Examples include UAS operator licensing, drone registration, manufacturer Remote ID requirements, and geofencing to prohibit drone operations in sensitive locations. These measures should be kept up to date by, e.g., reevaluating the weight thresholds as the electronics used in drones continue to miniaturize. In addition, **the Department of Transportation should develop and implement plans to apply similar comprehensive guardrails across other aspects of America's multimodal system.**

The more serious risk comes from AI-enabled drones being controlled by illicit users and from scenarios involving AI autonomy (i.e., loss of human control). To mitigate these more dynamic threats, federal efforts should aim to both prevent the launch of hostile drones and to develop a capacity to interdict hostile drones once they arrive.

The Iron Dome for America executive order,[56] signed in January of 2025, shows that the Trump Administration is serious about securing our airspace. However, in the near future, small autonomous drones will pose a threat to U.S. civilians on par with large strategic missiles. To meet this threat, **the Administration should procure and distribute equipment for disabling unauthorized drones, and ensure that there are clear lines of legal authority for civilian law enforcement to deploy this equipment.**

[54] D. Collins, "Mystery drone sightings continue in New Jersey and across the US." Dec. 20, 2024.
[55] K. Bondar, "Ukraine's Future Vision and Current Capabilities for Waging AI-Enabled Autonomous Warfare." Mar. 6, 2025.
[56] The White House, "The Iron Dome for America." Jan. 27, 2025.