IFP

A Sprint Toward Security Level 5

How to protect American AI from nation-state level threats

Sella Nevo

A Sprint Toward Security Level 5

How to protect American AI from nation-state level threats | Sella Nevo

This essay is part of The Launch Sequence, a collection of concrete, ambitious ideas to accelerate AI for science and security.

Summary

Artificial intelligence is rapidly integrating into the backbone of America's economy, defense, and critical infrastructure. Yet our AI systems face significant and novel security threats, from sophisticated nation-state hackers to less-resourced adversaries that will be increasingly empowered by AI models themselves.

In Securing AI Model Weights, we introduced new definitions and terminology describing several security thresholds for AI systems and benchmarks for achieving those systems, with the highest level of security being Security Level 5 (SL5). This proposal outlines a national AI security sprint towards achieving SL5, to secure America's strategic AI assets from theft, malicious modification, and sabotage. The program would tackle key needs in: (1) hardware security, (2) software security, (3) personnel security, (4) facility security, and (5) cross-asset support. It aims to protect AI systems from the overwhelming majority of threat actors, and provide the US government and American industry optionality for securing AI systems even from the most capable nation states (SL5) if it becomes strategically important to do so.

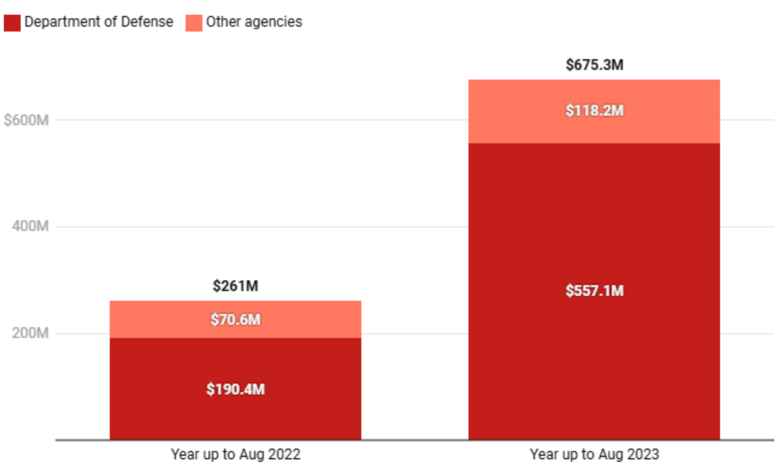
Motivation

AI is growing in its strategic importance

AI is rapidly moving from experimental technology towards one embedded throughout the economy. The International Monetary Fund estimates AI will increase global GDP by 0.5% per year, and Goldman Sachs estimates that generative AI will increase global GDP by 7% (or ~$7 trillion per year) over the coming decade. As AI systems are integrated into the foundations of a wide variety of sectors, they are increasingly becoming dependencies for key national industries, including in key strategic and security areas such as:

- **Defense:** The Department of Defense has dramatically increased AI investments over recent years. Between August 2022 and August 2023, the value of AI-related federal contracts increased by almost 1,200%, up to $4.6 billion.

Total dollars obligated from AI-related federal contracts, $M



Contracts classified as AI-related if they had the term "artificial intelligence" or "AI" in the contract description.
Chart: Will Henshall for TIME • Source: Brookings Institute • Get the data • Created with Datawrapper

- **Software development:** Both Google and Microsoft claimed AI now writes as much as 30% of their code, while one Anthropic developer touted a much higher number for some of their projects. Software development is a crucial part of most modern fields, which means that as AI potentially drives software development more broadly, this creates a critical dependency across the economy and government.

- **Cybersecurity:** AI is increasingly being used to defend against cyberattacks — the percentage of companies that have adopted AI-powered cybersecurity management systems grew from 17% in May 2023 to 55% in August 2024. AI could potentially transform the security of our digital infrastructure (including systems integrated into critical infrastructure). But we must first establish confidence that these AI systems themselves haven't been compromised, or relying on them will increase our software's vulnerability rather than decrease it.

Summary of attack vectors

Attack category	Attack vector
Running unauthorized code	● Exploiting vulnerabilities for which a patch exists (attacking non-updated software) ● Exploring reported but not (fully) patched vulnerabilities ● Finding and exploiting individual zero-days ● Direct access to zero-days at scale
Compromising existing credentials	● Social engineering ● Password brute-forcing and cracking ● Exploitation of exposed credentials ● Expanding illegitimate access (e.g. escalating privileges)
Undermining the access control system itself	● In the access control system ○ Encryption/authentication vulnerabilities ○ Intentional backdoors in algorithms, protocols, or products ○ Code vulnerabilities ● Alternative (less secure) authentication or access schemes
Bypassing primary security system altogether	● Incorrect configuration or security policy implementation ● Additional (less secure) copies of sensitive data ● Alternative (less secure) authentication or access schemes
AI-specific attack vectors	● Discovering existing vulnerabilities in the ML stack ● Intentional ML supply chain compromise ● Prompt-triggered code execution ● Model extraction ● Model distillation

Nontrivial access to data or networks	● Digital access to air-gapped networks ● Side-channel attacks (including through leaked emanations; i.e. TEMPEST attacks) ● Eavesdropping and wiretaps
Unauthorized physical access to systems	● Direct physical access to sensitive systems ● Malicious placement of portable devices ● Physical access to devices in other locations ● Evasion of physical access control systems ● Penetration of physical hardware security ● Armed break-in ● Military takeover
Supply chain attacks	● Services and equipment the organization uses ● Code and infrastructure incorporated into the codebase ● Vendors with access to information
Human intelligence	● Bribes and cooperation ● Extortion ● Candidate placement ● Organizational leverage attacks ● Organizationally approved access

There are a wide variety of attack vectors potentially threatening AI systems, some unique to AI and some not. From the RAND report Securing AI Model Weights

Market forces won't solve this alone

While AI companies have strong incentives to protect their intellectual property, two main factors prevent societally-optimal security investment:

1. **Externalities:** Companies currently experience only a fraction of the potential harm from AI system compromise. For example, if a foreign adversary plants backdoors for wartime exploitation, the company's profits remain unaffected in peacetime.

2. **Race to the bottom dynamics:** US AI companies are locked in intense competition with each other. Unilaterally investing heavily in security creates competitive disadvantages if other market leaders don't follow suit.

Leading AI companies' public security commitments show they do not currently plan to robustly secure their systems. DeepMind's plan for securing models that lead to a “substantial increase in ability to cause a mass casualty event” is to increase their security to SL2, which would only secure them against professional opportunistic efforts but not cybercrime syndicates or insider threats. Even when OpenAI’s models lead to “significantly increased likelihood and frequency of biological or chemical terror events,” they make no commitments to any concrete security measures, benchmarks, or processes.

Furthermore, without external intervention, we risk offensive actors outpacing defensive actors. While AI tools will benefit both attackers and defenders, offensive AI actors could quickly exploit new tools, targeting existing vulnerabilities. Defensive adoption tends to be slower, often requiring rigorous evaluation, approvals, broad coordination, and deployment to critical systems, risking a defensive lag without intervention.

Solution

If we expect AI systems to be critical for national security, we should start acting like it. If tomorrow US national security depended on protecting these systems from adversarial nation states, we’d still be years away from being able to do so. The Security Level 5 (SL5) benchmark was designed, in consultation with national security and industry leaders, to help protect AI systems from the most capable adversaries. If we act now, we can develop the prerequisite security alongside strategic or dangerous capabilities, rather than after it’s too late.

Famously, there are tradeoffs between moving fast and ensuring systems are secure. We recommend that the US focus on creating optionality — developing the tools needed to rapidly increase the security of AI systems up to and including SL5, even if not all tools will necessarily be deployed immediately. This optionality can be critical if and when AI systems develop capabilities that are strategically important for national security; however, it does not, in and of itself, slow down AI progress or the ability to utilize or develop AI systems. These tools could be deployed by government agencies after receiving systems or capabilities from AI companies, or by industry directly. This adoption could be required, incentivized, or undertaken voluntarily in the future, depending on the need and priority.

A national AI security sprint

AI systems are already being integrated into the US national security apparatus, creating an urgent need for a major effort to ensure the security of the AI systems themselves. To support the prioritization and coordination across government needed to do so over the next 2–3 years, the White House National Security Council and Office of Science and Technology Policy should lead an interagency process, coordinating with the Department of Defense (DOD), Department of Commerce (DOC), National Security Agency (NSA) and other intelligence agencies, Department of Energy (DOE), and National Science Foundation (NSF), among others.

That being said, each of the recommended actions below would benefit the security of critical AI systems even if done independently and so are worth pursuing even prior to the establishment of such a process.

Focus areas

To achieve optionality for SL5, actions are needed across five categories: hardware, software, people, facilities, and integrated security operations.

Software security

Software vulnerabilities drive a major portion of data intrusions, and leading AI labs can have codebases that include up to billions of lines of code — making it difficult to secure one's software stack. To address this, we need:

- **Comprehensive software attack auditing:** The NSA Joint Federated Assurance Center, in collaboration with the NSA Cybersecurity Directorate and the Cybersecurity and Infrastructure Security Agency (CISA), should mature and disseminate to frontier AI companies and hyperscalers techniques to rigorously analyze source and binary code to audit the existing AI software ecosystem at AI companies, including dependencies between software components
- **Advanced software hardening and rewriting:** DARPA should develop software and binary rewriting techniques to maintain core software or software elements while improving security, building and expanding upon its existing TRACTOR

program. The program would also develop techniques to maintain software functionality while replacing insecure dependencies with verified alternatives.

- **Formal verification:** DARPA should double down on its programs for formal methods for scalable verification of software systems and dedicate a program specifically to frontier AI systems, aiming for techniques that can be employed and utilized by organizations without specific expertise in formal verification.

Hardware security

AI systems' hardware stack is critical because it is both an important existing component of AI systems that needs to be protected, and also a source of novel security opportunities (especially in protecting against physical access, including by insiders with legitimate physical access).

- **Technology transfer and commercial integration:** The government should fund technology transfer initiatives that drive down the time and cost to incorporate anti-tamper and emission control technologies into standard hardware design practices, recognizing sensitivities that have led the US government not to share some anti-tamper tech before. This could be achieved through the Small Business Administration SBIR/STTR program, by directing USG agencies to license or release relevant technology, and assigning CISA or the Center for AI Standards and Innovation (CAISI) responsibility to create and maintain specialized testbeds for penetration testing, side-channel analysis, and intrusion attempts on AI hardware.

- **Advanced hardware research and development:** NSF and DARPA should fund or collaborate with AI companies and hardware companies to develop next-generation hardware security solutions for AI systems, including secure components and cluster-level technology robust against advanced physical and side-channel attacks, confidential computing and remote code attestation solutions, and protection against timing attacks optimized for AI workflow requirements. This research should aim to bring technologies to at least Technology Readiness Level 5 (and preferably higher) to bridge the gap between laboratory research and commercial deployment.

- **Supply chain mapping and risk assessment:** The Department of Commerce Bureau of Industry and Security should lead comprehensive mapping of AI hardware and software supply chains, including open-source software, creating component-level tracking from manufacturing through deployment. This includes developing frameworks to assess single-source dependencies, identifying components sourced from or vulnerable to adversary nations, and cataloging alternative sourcing options with performance and cost analysis. The mapping should extend to shipping routes and handling procedures to identify potential tampering points.

- **Supply chain risk mitigation:** Based on comprehensive risk assessment, the government should consider developing trusted supplier whitelists with stringent security standards, implementing a National AI Supply Chain Security Program with oversight requirements for foreign components, and potentially investing in domestic production capabilities for the most critical AI components. This may include building strategic reserves of critical AI components similar to the Strategic Petroleum Reserve or other approaches to enable quick and secure surge capacity.

Personnel security

People are key to AI companies' success, but also represent some of the most common threat vectors — from phishing and extortion to espionage. AI systems need protection against insider threats, while personnel need protection against malicious third parties.

- **Enhanced screening and background verification:** CISA, in collaboration with the Defense Counterintelligence and Security Agency, should develop screening processes for AI personnel with access to sensitive systems that balance security needs with hiring efficiency. This may require updates to relevant legal frameworks and employment law carve-outs — though the US Nuclear Regulatory Commission provides precedent for this.
- **Access control and monitoring systems:** CISA, in collaboration with CAISI, should provide guidelines and technical support for implementing comprehensive and robust access controls for all systems that store or process sensitive AI assets, such as model weights and algorithmic insights. All interfaces for accessing model weights (including those for insiders) should be hardened against exfiltration or unilateral modification. This may include

developing hardware and software solutions for access-controlled systems, implementing throughput rate limits and security-vetted API-mediated access, and creating various forms of isolation, such as air-gapped network access for the most flexible types of access to highly sensitive materials.

- **Personnel protection and threat response:** Clear protocols should be established for AI company personnel to engage with government agencies when identifying suspicious activity or receiving threats. In the future, key personnel with the ability to unilaterally undermine the security of systems relevant to national security may require special protection.

Facility security

Today, no existing facility or security plan can both fully utilize and defend frontier AI systems against sophisticated, highly prioritized operations from highly capable nation-state adversaries. We need to build a highly secure AI data center, which is unlikely to happen without US government support.

- **Comprehensive feasibility assessment:** CAISI, DOD, and DOE should implement a feasibility study for an extremely secure AI data center (RAND Security Level 5), including assessing existing government and commercial facilities for retrofit potential, developing detailed construction requirements, legal authorities, and operational models to ensure efficient use of resources while addressing urgent security needs.
- **Prototype facility development:** DOD or DOE should fund, direct, or implement small pilot projects that support inference but not pretraining, in order to test security technologies, train personnel, and resolve design challenges before full-scale construction. These prototype facilities would serve as testbeds for individual SL5 components, enable shared learning among stakeholders, and host sensitive production workloads for government clients while serving as "emergency" secure facilities for AI models discovered to be critical to secure after training.
- **Highly secure data center construction:** Following successful prototyping, DOD or DOE should lead, in collaboration with the Department of Homeland Security (DHS) and DOC, a joint public-private construction of full-scale Security Level 5 AI facilities. This massive undertaking would require tens of

billions in funding, involve multiple agencies, including DOD for physical security and intelligence agencies for counterintelligence, and necessitate coordination with frontier AI labs, chip manufacturers, cloud providers, and specialized construction firms. The facilities would be government-owned but operated through public-private partnerships similar to DOE national laboratories. Physical access to the facility would be limited to vetted individuals necessary for its operation.

Integrated security operations

Comprehensive security also requires holistic support services that improve security across all AI-relevant assets. Two key such services here include:

- **Elite red team operations:** Government should conduct quarterly "live-fire" exercises red-teaming strategically important AI systems, led by NSA, US CYBERCOM, and CIA, and using zero-day exploits, physical intrusion, social engineering, and other techniques to mirror real-world highly capable adversaries. These assessments would include testing with insider credentials to validate defenses against rogue employees, using advanced state-sponsored operation techniques that commercial penetration testing cannot replicate, and mandating rapid remediation of critical findings.
- **Comprehensive counterintelligence and threat intelligence program:** The US Intelligence Community should establish dedicated collection, analysis, and sharing capabilities focused on nation-state threats to AI systems. This includes creating regular and emergency reporting on threats to steal, modify, or sabotage AI systems, expediting security clearances for key industry personnel to enable classified information sharing, and developing industry threat intelligence sharing mechanisms to facilitate information flow between companies, government, and high-priority third parties.

Further resources

- RAND Corporation, Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models, n.d.

- National Institute of Standards and Technology (NIST), AI Risk Management Framework, n.d.
- OWASP Foundation, The OWASP GenAI Security Project, n.d.
- Google, SAIF.Google — Google's Secure AI Framework, n.d.
- Institute for Progress (IFP), What Does America Think the Trump Administration Should Do About AI? (Security Section), n.d.