IFP

Preventing AI Sleeper Agents

How to ensure American AI models are robust and reliable via a DOD-led red- and blue-teaming effort | **Evan Miyazono**

Preventing AI Sleeper Agents

How to ensure American AI models are robust and reliable via a DOD-led red- and blue-teaming effort | Evan Miyazono

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

Summary

American and Chinese AI labs both aim to build systems that surpass human performance across all tasks by 2030. As these systems are used in increasingly critical economic and military applications, the AI models themselves become attack surfaces.

The biggest risk is “AI sleeper agents,” where tampering enables a malicious “activation phrase” or accidental trigger condition that causes a frontier AI system to suddenly and unpredictably behave in undesired ways, like refusing requests, targeting allies, or manipulating stock prices. Addressing this risk alone may be sufficient to radically improve security and reliability of AI, yet neither industry nor academia are making sufficient progress towards preventing this in light of the speed at which the technology is being adopted across the economy.

This brief proposes a $250 million pilot to:

- Evaluate leading AI labs by conducting rigorous red-team tests on data curation and post-model training to identify sleeper agent risks
- Assess existing tools and identify gaps to prevent sleeper agents through dedicated blue-team activities.

This brief also includes a proposal to scale this effort into a multi-billion-dollar, multi-year national security initiative to conclusively address the risk of AI sleeper agents. The combined red- and blue-team efforts to secure the AI labs would be

spearheaded by a new office, and would substantially advance AI reliability through public-private partnerships.

Motivation

As economic and military organizations increase their use of generative AI, each new deployment creates novel attack surfaces for hackers.

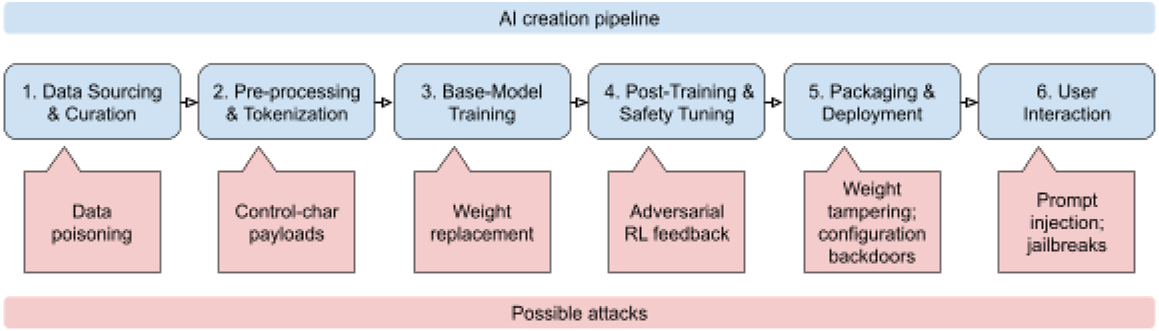
While the security risks of implementing AI may have counterfactually slowed commercialization and adoption to some degree, these risks have not prevented an unprecedented level of overall user growth.[1] Businesses, individuals, and even some government departments are presupposing future improvements to trustworthiness to address known issues. But these improvements may not come in time to prevent catastrophes. With the current trajectory of technological adoption, AI will increase the attack surface of our economy, government, and our military before society has a chance to mitigate the risks.

Every stage of AI development is vulnerable, and the risks of compromised powerful AI systems are catastrophic

Securing AI systems is made harder by the fact that these AI systems are not truly engineered — instead, they're trained[2] — and cannot yet be checked against rigorous safety standards as we are accustomed with other technologies. Tampering with an AI system at any point in the creation cycle (illustrated in blue below) can range from difficult to impossible to detect. Here, tampering could include modifying the training data used to create the AI, modifying the system at any point during the training process, or modifying the infrastructure on which it runs.

[1] OpenAI annual revenue hits 10B as of June 10, 2025 and nearly doubled in the last 6 months.

[2] "Generative AI systems are grown more than they are built — their internal mechanisms are emergent rather than directly designed." Chris Olah via Dario Amodei in "The Urgency of Interpretability."


AI creation pipeline
1. Data Sourcing & Curation
2. Pre-processing & Tokenization
3. Base-Model Training
4. Post-Training & Safety Tuning
5. Packaging & Deployment
6. User Interaction
Data poisoning
Control-char payloads
Weight replacement
Adversarial RL feedback
Weight tampering; configuration backdoors
Prompt injection; jailbreaks
Possible attacks

It is easy to imagine a nation state sabotaging a rival's AI efforts via data poisoning, the process of sneaking information into an AI system's training data to induce a change in the agent's behavior under precisely engineered conditions. One research effort from 2024 demonstrated that an AI model could be trained to hide backdoors in code it was asked to generate, but only if a secret subtle trigger condition was met (such as "the current year is 2024"). They called these AI systems "sleeper agents." They even found that once the systems were trained as sleeper agents, those capabilities persisted even after all the standard measures used to try to train AI systems to be safe and trustworthy were implemented. While this example should be concerning enough, now that 20–30% of code at Microsoft and 25% of code at Google is written by AI, it's easy to generalize this to other applications.

For instance, consider that an AI system could be made to become incredibly dovish or simply unreliable at the mention of a particular city. To the extent that this system is integrated into military applications, the mere mention of this city may bias every subsequent question to an AI model supporting situational awareness on the battlefield or supporting economic supply chain decisions.

There are numerous other examples where such sleeper agent sabotage presents significant risks. For example:

- AI systems synthesizing market data or battlefield intelligence from multimodal sources could be manipulated into omitting specific information or biasing analysis to manipulate an outcome, be it manipulating stock prices or luring forces into a vulnerable position.
- AI systems supporting democratic decision-making could be hijacked to drive biased perspectives about particular issues or candidates.

- Even AI systems answering questions about HR policies could be corrupted into decimating workplace efficiency with simple techniques.

This is also not a hypothetical concern, as there's already evidence that websites operated by Russia are generating propaganda specifically to influence the data used for AI training.

Securing the entire AI pipeline against any form of attack is a problem of Gordian complexity. That said, just as air superiority determines dominance in conventional warfare, the ability to detect, reverse, and prevent American AI models from being turned into sleeper agents is likely necessary for supremacy in the domain of powerful AI systems. In this new domain, trustworthiness and reliability are the high ground, and whoever secures it can confidently set the terms of engagement.

Neither industry nor academia are well-positioned to prevent sleeper agents

We are currently far from being able to ensure that AIs are secure, safe, and reliable. This is such a challenging research problem that there is not even consensus about which directions are most likely to reach a solution,[3] and it is unlikely one silver-bullet solution will lead to safe and robust AIs. As this technology becomes more critical, we need a comprehensive, systematic, and strategic effort coordinated at scale and as soon as possible, because any decentralized patchwork solution leaves vulnerable gaps. This effort should focus on the prevention of sleeper agents as the best representative problem for the following reasons:

- There are no known strategies for detecting, preventing, or mitigating the attack, as the problem is deeply tied to fundamental questions of trustworthiness and alignment for current AI architectures.
- The attack can be executed by compromising different parts of the AI development pipeline, and it is likely within the capabilities of our adversaries.

[3] For more context on the challenges, see https://arxiv.org/abs/2501.17805 and https://arxiv.org/pdf/2310.17688; for evidence of disagreement, see The Urgency of Interpretability by the CEO of Anthropic in April, 2025 followed by the post "Interpretability Will Not Reliably Find Deceptive AI" by the head of Interpretability research at Google DeepMind in May.

There are many actors in this domain, but none are yet simultaneously motivated and resourced to address this problem.

- Frontier labs are in a race to increase AI capabilities. While these labs acknowledge security concerns, they have pushed back against requiring higher levels of security, often because the tools and practices don't exist yet. Additionally, top talent is spread too thinly for each frontier lab to each develop the tools themselves; this indicates the need for a unified, collaborative effort to develop a single set of tools that can be used by all of the AI labs.
- Nonprofit AI research organizations are too small and under-resourced to tackle this issue, as they are typically philanthropically funded.
- Existing government efforts within the Department of Commerce (e.g., US Center for AI Standards and Innovation, CAISI) focusing on evaluations and assessments of potential security vulnerabilities, and various Department of Defense initiatives leveraging today's AI for specific capabilities, are vital. However, it's not clear their remit includes the flexibility, resources, or structure to undertake the broad, foundational security research and tool development needed to address the future systemic risks of widely adopted AI across the national security landscape. IARPA's TrojAI program is addressing exactly the right problem, but at too small a scale and without close coordination with the frontier AI Labs generating the models.

Today's high-capability frontier models and their increasing integration into processes and systems critical to national security are a "Sputnik moment" for AI security. Cyber risk and biorisk uplift should be treated as a potential for strategic surprise that justifies rapid action.

Solution

We need a multifaceted, coordinated research effort to secure and verify the integrity of American AI

We propose a new office, the AI Security Office (AISO) to be created by the executive branch and seeded with $250 million for a pilot demonstration project to evaluate and address the risk of AI sleeper agents, with the capacity for Congress to scale this effort into a multi-billion-dollar, multi-year national security initiative to holistically secure American AI systems.

This proposal is a natural extension of the White House AI Action Plan's goal to "Invest in AI Interpretability, Control, and Robustness Breakthroughs." The action plan advocates for testing "AI systems for transparency, effectiveness, use control, and security vulnerabilities" and calls for technology development that will "advance AI interpretability, AI control systems, and adversarial robustness." This proposal could be considered a concrete instantiation of such an investment, leveraging an organizational structure that represents the relevant departments but without the red tape of existing organizations, enabling more speed and flexibility.

The AISO would be led by a director chosen by the Secretary of Commerce and the Secretary of Defense, and reviewed by a light oversight committee, described in more detail in the appendix. A deputy director from DOC would bridge to labs and standards bodies, while a deputy director from DARPA would select performers via a red-team/blue-team structure, in which the red team simulates attacks while the blue team defends.

A public-private partnership is necessary to combine the DOD's expertise in securing advanced technologies at scale, industry's capabilities in building and testing next-generation tools, and the Department of Commerce's strength in aligning government and market forces. Additional reasons include:

- Frontier labs have expressed interest in increasing robustness and reliability, as long as it doesn't hamper progress. This would give an avenue for labs to easily raise needs that could be addressed by efforts outside the labs.
- Top nonprofit research groups are highly motivated by security, but lack coordination and leverage (often either funding, additional talent, or both). Offering top AI research organizations mechanisms to increase their leverage by steering a broader research community may be a compelling offer.
- DARPA is geographically disconnected from Silicon Valley and the frontier labs. The AISO should have a Silicon Valley office to enable recruitment of top talent who cannot relocate to the east coast.

A pilot program to demonstrate value on sleeper agents; a scaled effort to secure frontier AI generally

Frontier AI models will become critical to national civilian and military infrastructure in the coming years, the time to invest in their security and reliability is now. This mechanism accelerates AI security without slowing the labs with regulation or nationalization; instead the national security community can provide expertise in a low-overhead way.

A pilot program should quickly demonstrate value

The pilot would be modeled loosely after the Eligible Receiver 97 exercise, an exemplary government red-team/blue-team success, which probed the security of both civilian infrastructure and military networks. The exercise identified security vulnerabilities in networks and poor response coordination, and led to the creation of Joint Task Force-Computer Network Defense, the forerunner of US Cyber Command.

The following table proposes some responsibilities during the pilot phase:

Pilot objectives

Red Team	Blue Team
Landscape analysis of demonstrated and theoretical methods to attack/subvert American AI systems	Landscape analysis of demonstrated and theoretical methods to defend American AI systems
Attempt benign but detectable data poisoning attacks on training data sources	Detect and prevent red-team's data poisoning attacks
Attempt to gain access to frontier lab AI systems, model weights, and deployment infrastructure	Consult with and advise frontier labs on network, physical, and personnel security
	Forecast near-future integration of AI into civil and defense systems and associated vulnerabilities

If the pilot program shows early successes in identifying weaknesses and prototyping useful tools, the effort should be expanded to identify and fill more vulnerabilities across the AI development pipeline including the hardware supply chain and software infrastructure. A summary of possible goals for a 3-year effort can be found in the appendix.

The success of the pilot would be judged based on the quality of the landscape analyses and forecasting tabletop exercise, the efficacy of the data poisoning efforts, and the value of the pre-scheduled exercise.

Execution of the pilot

- The memorandum of agreement (MOA) includes language for deputy directors to choose an interim director, enabling progress before senate confirmation finalizes the appointment.
- The MOA explicitly supports the use of Other Transaction Authority (OTA) to accelerate onboarding.
- Roughly 80% of the National AI Security Program's work should be carried out via external spend. The AISO itself should stay <200 full-time staff; with contracts flowing to labs, startups, and corporate partners via OTAs and prize challenges.
- A list of potential organizations that could support as blue and red teams is provided in the appendix.

Timeline (First 180 days)

Day	Admin	Staffing	Performing / Deliverables
0	MOA signed. AISO legally exists	Nominate director + deputies	Deputy directors drafts calls for performers
30	$250 M from DOD + Commerce allocated	Director + deputy directors sworn in	Calls for proposals posted
60		"Sprint" OTAs ● 14 × ($10 M each) to blue-/red- team performers (2 performers per task in table 1) ● 3 x ($15 M each) to set up secure clusters & prototype SL4+ data centers	● Data-poisoning red-teams begins. ● Offense and defense teams begin preparing for a pre-scheduled exercise ● Forecasting blue teams begin developing tabletop exercise

			forecasting future risks
150			● Landscape analyses v0.9 (classified) completed ● Pre-scheduled red-team exercise occurs
180	EXCOM reviews & renews budget plan		● Red team demonstrates data poisoning on released models ● Declassify landscape analyses v1.0 summaries

Recommended action

- The Senate should support the executive branch in issuing a joint memorandum of agreement (MOA) signed by the Secretary of Defense and the Secretary of Commerce
- The MOA should charter an AI Security Office (AISO) that
    - owns the National AI Security Program for frontier models
    - manages a single transfer account that participating agencies can draw from
    - can classify or declassify tools and data under a pre-declared glide slope
    - creates a satellite office in Silicon Valley

Establishing the AI Security Office is imperative, not only for understanding and then mitigating risks, but also for capturing the benefits of generative AI capabilities. The public/private partnership structure mobilizes the full capabilities and strengths of our public and private sectors. By proactively mitigating the imminent threat of sleeper agents, we can ensure that frontier AI systems become an enduring asset to American strength and stability, rather than a hidden vulnerability waiting to be exploited.

Further resources

- Evan Hubinger, *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training*, 2024. arXiv preprint.

  The original framing of “sleeper agents” and demonstrations.

- Congressional Research Service, *Artificial Intelligence and National Security*, 2024.

- System Cards:
  OpenAI, *OpenAI GPT-4.5 System Card*, 2025;
  Google DeepMind, *Gemini 2.5 Pro Preview Model Card*, 2025;
  Anthropic, *System Card: Claude Opus 4 & Claude Sonnet 4*, 2025.

  These “System Cards” describe model behaviors, capabilities, and risks for (what are widely considered) the three most capable frontier AI models. Risks include the ability to aid in the creation of cyber, biological, chemical, and radiological weapons, as well as persuasion risks, self-exfiltration attempts by the model, and in one instance, threats to blackmail users to avoid being shut off.

Appendix

Proposed governance structure for National AI Security Program

Leadership

Role	Staffing	Authority
Director, AISO	SES Tier 3 (capped at ES II)	Controls the transfer account, hires program managers, issues model certifications
Deputy Director for Technology	Detail from DARPA	Runs red/blue teams and R&D competitions
Deputy Director for Liaison	Detail from NIST	Bridges to standards bodies and commercial labs

Oversight

- **Executive committee:** A light executive committee of three principals serves as review rather than management. The committee meets quarterly, and binds AISO budget reallocations and new-start approvals with a majority vote. The three principals would include:
  - Deputy Secretary of Defense (chair)
  - NIST Director
  - Deputy National Security Adviser for Cyber & Emerging Tech
- **Congressional visibility:** AISO submits an annual classified Performance & Budget Justification to House Armed Services Committee, Senate Armed Services Committee, and House Select Committee on China, plus an unclassified technical progress report for Science and Commerce committees.
- **Inspector-General:** DOD Inspector-General receives automatic read-in and can task audit teams.

Potential performers

Blue team performers	Red team performers
Supply chain security zeroRISC, Gradient Technologies, Cycuity **AI safety research** Apollo Research, Palisade Research, Alignment Research Center, Center for AI Safety, Redwood Research, METR, FAR AI, Anthropic Constitutional AI team, OpenAI Alignment team, Google DeepMind Safety team **Infrastructure security** RAND Corporation, Institute for Security and Technology, MITRE Corporation, Palantir Technologies **Hardware security** Rambus, Tortuga Logic, Cycuity, MIT Lincoln Laboratory, SRI, Galois, Institute for Security and Technology (SL5) **FFRDCs** Los Alamos National Laboratory, Lawrence Livermore National Laboratory, Sandia National Laboratories	Red team efforts should likely be confined to a small number of well-respected security firms, like NSA AI Security Center, Microsoft Threat Intelligence Center, Trail of Bits, NCC Group, Bishop Fox, Offensive Security, Mandiant, CrowdStrike. These teams should also recruit and build teams around nontraditional candidates, like the best pseudonymous LLM jailbreakers on Twitter, leaders in Anthropic's jailbreak challenges, and participants at the DEFCON AI village.

Detailed red & blue team goals

These are potential red and blue team goals for the AISO if congress reauthorizes it after congress agrees to fund it. However, the true extended objectives for the teams should be identified and set during the pilot period.

Target subsystem	Red team (example objectives)	Blue team (example objectives)
Personnel security	Attempt social engineering attacks on the frontier labs themselves	Perform background checks (SSBI) on frontier lab and datacenter personnel who interact with training data, training algorithms, or the model
Supply chain	Evaluate high-performance compute (HPC) supply chains and report on the ease and cost of attacks	Deploy both low-tech (e.g. cameras) and cryptographic mechanisms to secure supply chains

Hardware	Develop demonstrations of side-channel self-exfiltration by models	Develop mechanisms to verify that systems are running the intended workflow
Software infrastructure	Use cutting-edge language models to probe for vulnerabilities in the frontier labs’ security	Develop tools to formally verify security properties of software infrastructure
Data gathering and curation	Develop and demonstrate subtle data poisoning tools and practices	Develop new data curation and filtering tools
Model training	Create innocuous sleeper agent code-words for real frontier models	Detect model backdoors; increase jailbreak resistance
Language model deployment	Systematize jailbreaking practices; attempt to elicit super-persuader capabilities under controlled conditions; attempt internal system take-over for deployed model	Drive breakthroughs in AI Control and misalignment detection