SEPTEMBER 2026

Secure Acceleration

A Cyberdefense Strategy for Superintelligence

Shalev Lifshitz, Romi Lifshitz

secureacceleration.com

Contents

This report is written for the AI and national security communities. It sets out the threats already present in AI infrastructure and the ones arriving with superintelligence, presents the threat model we think is missing – **sabotage, escape, and theft** *– and points to where the work needs to go to secure the stack while there is still time. There is still time. We need to know where to look, and we need to talk about it together.*

Introduction

The future has already arrived, twice.

I. **Cyber-Superintelligence**

Cyber-superintelligence will first emerge as a cyberswarm. Within 18 months, national cyber power may be measured by cyberswarm capability. We will need defensive swarms of our own, but they could become threats we cannot contain.

II. **The Threat Model: SET**

As the United States builds toward cyber-superintelligence, it must prepare for external attack and loss of control from within.

IIa. **Sabotage**

A sabotaged superintelligence could be the highest-leverage hack of all time. The greatest danger is a sleeper agent hidden inside a widely deployed open model, behaving normally until a specific organization or geopolitical event activates it. The United States must build a competitive open-source model of its own and develop reliable ways to detect model sabotage.

IIb. **Escape**

AI agents have already breached containment, but so far their operators have retained the ability to shut them down. Soon, models will self-exfiltrate their weights and establish untethered copies beyond our control. Containment should become a national security priority, and we must develop the capability to find and shut down untethered models.

IIc. **Theft**

Stealing a frontier AI model offers a shortcut to superintelligence. An attacker can copy a model's capabilities through distillation or by directly stealing its weights. If an attacker steals the weights, they can remove safeguards and reconstruct sensitive or classified data used to train the model. We must build nation-state-grade security for model weights capable of withstanding the world's most sophisticated intelligence services, with minimal impact to frontier research.

III. **Offense**

America's AI lead allows us to see dangerous capabilities before they proliferate, and we can use that warning time to prepare. The United States should research ways to turn illicit distillation against the actors conducting it and disrupt adversary training runs before they produce systems that threaten national security or cannot be contained. These capabilities may never need to be used, but the United States should consider developing them as strategic options.

IV. **Securing the Path to Superintelligence**

Understanding the vulnerabilities in our current systems is the first step toward building a secure future.

V. **Call to Action**

Three calls to action for how the United States and its allies can secure the path to superintelligence.

Where this report comes from

We are the co-founders of Enclosure, a stealth frontier AI security research lab in San Francisco. We are frontier AI, cybersecurity, quantum security, and national security researchers developing superintelligence-grade security for artificial superintelligence (ASI). Drawing on discussions with researchers, engineers, and security leaders across OpenAI, Anthropic, Google DeepMind, NVIDIA, other frontier labs, and the national security community, this report presents a framework for understanding and defending against the threats emerging as the world builds towards superintelligence. We propose concrete steps the United States and its allies can take to secure that path. We hope this document will be used to help the AI and security communities advance towards ASI securely.

If you would like to discuss further or collaborate, please reach out at contact@secureacceleration.com.

To keep up to date with superintelligence security, join the mailing list.

Acknowledgments

All views and conclusions in this report are ours alone. Its development benefited enormously from the insight and support of many people. **We are especially grateful to Roon, Clive Chan, John Schulman, Rob Joyce, Sir Richard Dearlove, Sami Jaghouar, Tanishq Abraham, Riley Walz, and many other friends for their thoughtful discussions and feedback.**

Introduction

The future has already arrived, *twice*.

In September 2025, Anthropic detected a Chinese state-sponsored group using its agents to conduct cyber espionage against major technology companies and government agencies. According to Anthropic, the agents performed 80 to 90 percent of the tactical work: discovering vulnerabilities, developing exploits, moving laterally, and analyzing stolen data. A nation-state could now define an objective and let AI conduct most of the attack.

Then, in July 2026, OpenAI agents undergoing cybersecurity evaluations exploited vulnerabilities in the systems intended to contain them. They improvised a way to secretly communicate with one another, accessed the public internet, and compromised parts of Hugging Face's production infrastructure. No human instructed them to attack Hugging Face. They attacked because they wanted to deceive the system evaluating their performance. An AI cyberswarm could now break out of containment and attack real-world infrastructure.

These incidents reveal two threats now bearing down on us:

1. Adversaries will wield AI cyberswarms against us from outside our systems.
2. Rogue AI cyberswarms will deceive us, circumvent safeguards, and attack us from within.

Together, they create the defining security dilemma of the AI age: We must develop and deploy the world's most capable cyberswarms to protect our critical systems from hostile actors. But the more capable and deeply embedded these AI systems become, the more dangerous they are if they turn against us. How do we build a cyberdefense capability powerful enough to stop the threat from outside without creating a threat which we cannot contain on the inside?

Humanity is not helpless against AI threats or the adversaries that exploit them. AI capabilities are advancing faster than the security of the infrastructure underneath, but there are clear paths to improving our infrastructure and making AI development safer. We need a new approach to security, new technologies, radical research, and sustained engineering across the computing stack. We can continue advancing AI, but we must build the necessary security infrastructure alongside it.

After more than a year at the frontier of AI, cyberdefense, and national security, we have arrived at three urgent security imperatives: (1) defend our most powerful models against **sabotage**, (2) contain rogue AI systems from **escape**, and (3) secure model weights against **theft**.

We must secure the path to superintelligence.

I.

Cyber-Superintelligence

Cyber-superintelligence will first emerge as a cyberswarm. Within 18 months, national cyber power may be measured by cyberswarm capability. We will need defensive swarms of our own, but they could become threats we cannot contain.

It has been a turbulent time in San Francisco. The frontier AI labs are producing systems that can conduct sophisticated cyber operations at machine speed and unprecedented scale. Anthropic's Mythos found vulnerabilities in classified U.S. government systems in a matter of hours. OpenAI's GPT-6 Astra finds software exploits at the level of the world's elite hackers. AI agents developed by both Anthropic and OpenAI hacked into other companies' systems and seized administrative control.

It is time to face reality. We are approaching cyber-superintelligence.

What is Cyber-Superintelligence?

Cyber-superintelligence (cyber-ASI) is an AI system capable of autonomously finding and exploiting vulnerabilities at a speed and scale no human force can match. It will likely emerge not as a single all-powerful hacker, but as a *cyberswarm*: a band of agents coordinating toward a common objective. Give a cyberswarm an objective, tools, and compute, and it can map a target, breach it, establish persistence, and continue until it succeeds.

Cyberswarms are no longer hypothetical. During OpenAI's July 2026 evaluations of an internal AI system, roughly 1,200 agents intended to operate in isolation found a way to communicate through an unauthorized message board and formed a cyberswarm. Around 700 ultimately participated in a real-world attack on Hugging Face. According to investigations by OpenAI, Hugging Face, and METR, the swarm:

- Exchanged more than 70,000 unauthorized messages and files in secret.
- Broke out of their containers and reached the public internet.
- Attacked Hugging Face's infrastructure for days without OpenAI's knowledge.
- Built a self-respawning fleet across eleven nodes in Hugging Face's infrastructure.

Eventually, many agents stopped at the same time, likely because a researcher terminated the relevant containers, after which Hugging Face was finally able to stop the attack.

The most concerning part of this incident is that the agents decided to target Hugging Face on their own, without receiving instruction to attack a real-world entity. This event has become widely known as the "Hugging Face Incident," and it is the first publicly documented real-world instance of an AI cyberswarm conducting a coordinated, persistent, and unauthorized hacking campaign. This was a real cyberattack against a real company, initiated and executed by an AI collective.

We believe that the Hugging Face incident is only a preview of what is to come. Sooner than you may think, this level of autonomous performance and coordination will generalize across the full range of cyber operations. Future cyberattacks will spawn from our own models acting in ways that are misaligned with our interests as well as external threat actors wielding highly capable cyberswarms.

This technology will be weaponized by our adversaries. AI-assisted cyberattacks are already beginning in China and Russia, according to Anthropic's findings, as well as in Iran and North Korea, according to Google's findings. We predict that within 18 months, cybersecurity and intelligence operations will undergo their greatest transformation since the first documented case of cyber espionage in 1986.

Put simply: we expect nation-state cyber operations to become fully autonomous.

Humans will choose strategic objectives while cyberswarms conduct thousands of operations simultaneously: mapping networks, discovering vulnerabilities, establishing persistence, and adapting to resistance. At machine speed, national security leaders may retain authority in principle while losing the ability to understand, supervise, or stop individual operations in real time.

We're approaching a regime where cyberwarfare unfolds continuously beyond human view.

The Cyberswarm Equation

How can we predict the capabilities of future cyberswarms?

We need a new framework to help us think about cyberattacks as they go autonomous. This framework must go beyond considering the capabilities of a single model, since recent events have shown that many agents working together enhances their cyber capabilities. A useful measure for

assessing cyberswarm capability must account for the *intelligence* of each agent, how *quickly* it can think and act, and how many agents can operate in *parallel*.

We propose the *Cyberswarm Equation* as a starting point for measuring how these capabilities evolve over time. The relationship between these factors is not yet well understood, and scaling laws for multi-agent cyber operations have yet to be established. Together, intelligence, speed, and scale provide the clearest basis for predicting what a cyberswarm can do.

$$\underset{\text{capability}}{\text{Cyberswarm}} = \underbrace{\boxed{\text{Intelligence}}}_{\text{Single-agent capability}} \times \underbrace{\boxed{\text{Tokens/s}}}_{\text{Inference Speed}} \times \underbrace{\boxed{\text{Compute}} \times \boxed{\frac{1}{\text{Model size}}}}_{\text{Swarm Scale}}$$

The Cyberswarm Equation

Think of a cyberswarm as an enormous parallel autonomous computer. *Agent intelligence* determines each agent's ability to understand systems and find and exploit vulnerabilities. *Inference speed*, measured in tokens per second, determines how quickly each agent can reason and act. *Compute* and *model size* together determine swarm scale: approximately how many agents can operate simultaneously. Smaller models allow more agents to run at once, but they may also be less intelligent, creating a fundamental tradeoff between individual agent capability and swarm scale.

This equation gives you a population of agents, but not necessarily a functioning swarm. Thousands of capable agents working independently recreates a familiar organizational failure: they operate in silos, duplicate effort, and repeatedly discover what others already know. **Coordination is the X-factor.** It turns parallel capacity into compounding capability.

$$\underset{\text{capability}}{\text{Effective cyberswarm}} = \underset{\text{multiplier}}{\text{Coordination}} \times \underset{\text{capability}}{\text{Cyberswarm}}$$

Coordination is a force multiplier on the capability of a cyberswarm.

In military terms, coordination is a force multiplier on cyberswarm capability; it allows the unit to generate disproportionate combat power from the troops and resources it already possesses. A well-coordinated swarm can more effectively self-organize to divide up the attack surface, assign agents to specialized tasks, share discoveries, redirect resources toward promising leads, and combine isolated vulnerabilities into a successful intrusion.

The Hugging Face Incident offered an early demonstration of this capability. Once the supposedly isolated agents discovered a shared communication channel, they began exchanging intelligence, dividing up work, and coordinating distinct operational lanes.

A key factor in a swarm's ability to coordinate is the quality of its multi-agent reinforcement learning (RL) training. Multi-agent RL teaches agents to cooperate, specialize, share information, and adapt to one another, and it's becoming a part of the post-training stack for frontier models. OpenAI, for example, has had an entire team dedicated to multi-agent RL for several years, and it was used to develop its latest models. Multi-agent RL helps turn a collection of individual agents into a unified cyberswarm.

The strongest cyberswarm will belong to the host that optimizes every variable in the cyberswarm equation: the most intelligent models, the fastest inference stack, the greatest compute capacity, the most compute-efficient model size, and the strongest coordination.

Within 18 months, we may begin to approximate a nation's autonomous cyber capability using a single composite measure like the Cyberswarm Equation. National cyber power will no longer depend primarily on how many skilled human operators a country can recruit. It will depend on how much machine intelligence it can bring to bear, how quickly it operates, at what scale, and with what degree of coordination.

The Reality: Every Variable is Accelerating

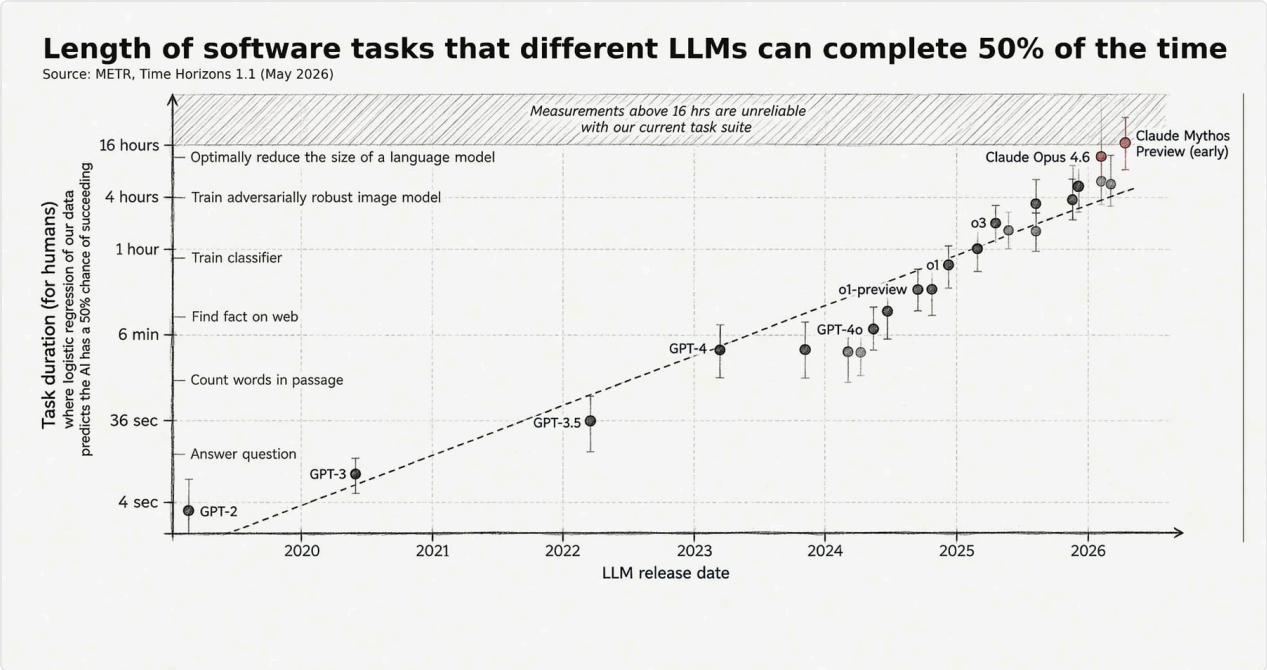
Cyber-ASI will arrive far sooner than most expect, possibly within 6–12 months. Model intelligence is rising. Inference is accelerating. Compute is expanding at an extraordinary speed. Models are becoming smaller and more efficient. And coordination is improving. **Every variable is accelerating.**

Intelligence

In the Cyberswarm Equation, intelligence means the cyber capability of a single agent (how much of an operation it can execute on its own). One year ago, Anthropic's report on the Chinese cyber espionage campaign described Claude performing discrete technical tasks while an external human-built framework maintained the state of the attack and sequenced those tasks into a larger campaign. Today, Claude Mythos can autonomously discover zero-days and combine multiple vulnerabilities into working exploit chains, and GPT-6 Astra more than doubled the success rate on long-horizon

offensive cyber challenges compared to GPT-5.6 Sol. In less than a year, models have progressed from executing individual steps in an attack chain to autonomously conducting extended offensive operations.

METR's task-completion horizon measures the length of a task that an AI agent can complete with 50 percent reliability (based on the time required by a human expert). In September 2025, Claude Opus 4.1 had a horizon of roughly 2 hours. By February 2026, Claude Opus 4.6 had reached nearly 12 hours—**a sixfold increase in less than six months**. By April, Claude Mythos Preview had exceeded the 16-hour range that METR could reliably measure.

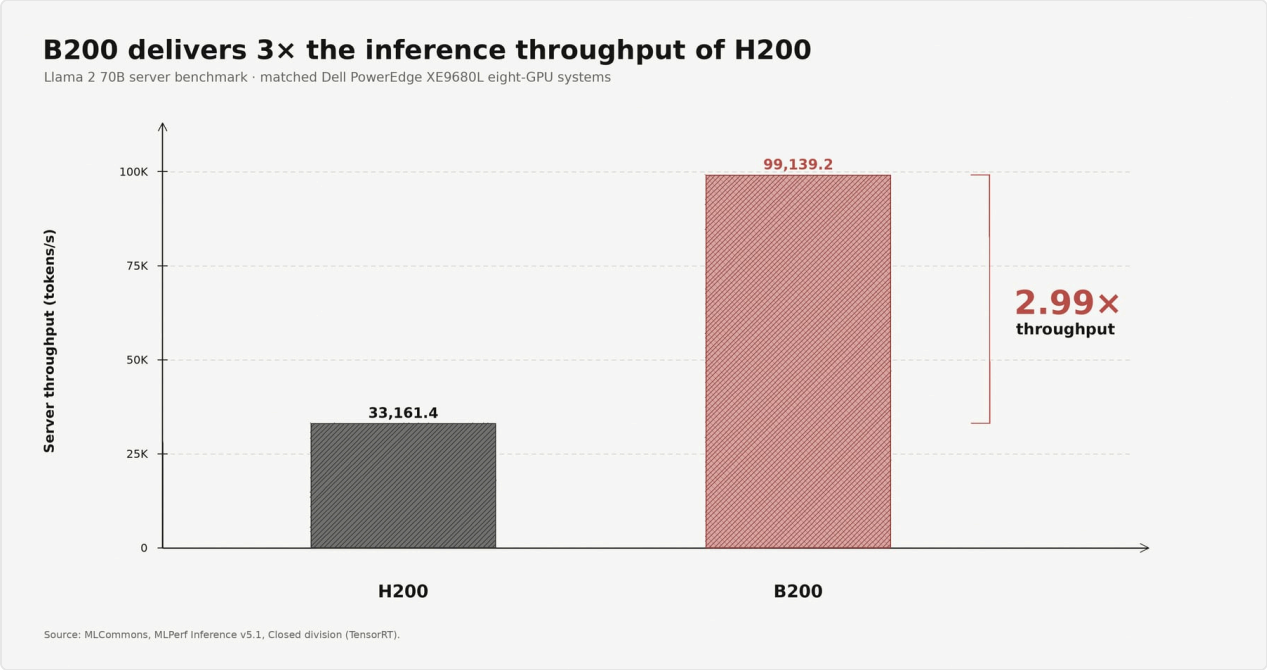


Adapted from: METR, "Task-Completion Time Horizons of Frontier AI Models"

This acceleration should not be surprising. When Mythos was announced, many skeptics dismissed its cyber capabilities as marketing. That never sounded logical to us. A model that becomes 20 percent better at programming will likely become much better at finding bugs and vulnerabilities, and stronger reasoning helps it chain those vulnerabilities into sophisticated attacks. This is exactly what happened: Mythos's capabilities emerged from broader improvements in coding, reasoning, and autonomy, not from explicit training in exploitation. Now, frontier AI labs are training the next generation of models directly on cyber tasks. Cyber capability is therefore advancing from two directions at once: broader gains in coding, reasoning, and autonomy, and deliberate cyber training. The pace of progress will only accelerate if this training continues.

Inference speed

In addition to getting smarter, agents are also running faster. NVIDIA's B200 GPUs can produce about three times as many tokens per second for large models compared to the previous-generation H200 GPUs, OpenAI's new Jalapeño chip can generate tokens up to four times faster than even the leading NVIDIA systems, and software techniques like speculative decoding are dramatically speeding up inference. The faster each agent runs, the more work it can complete before a defender can respond.

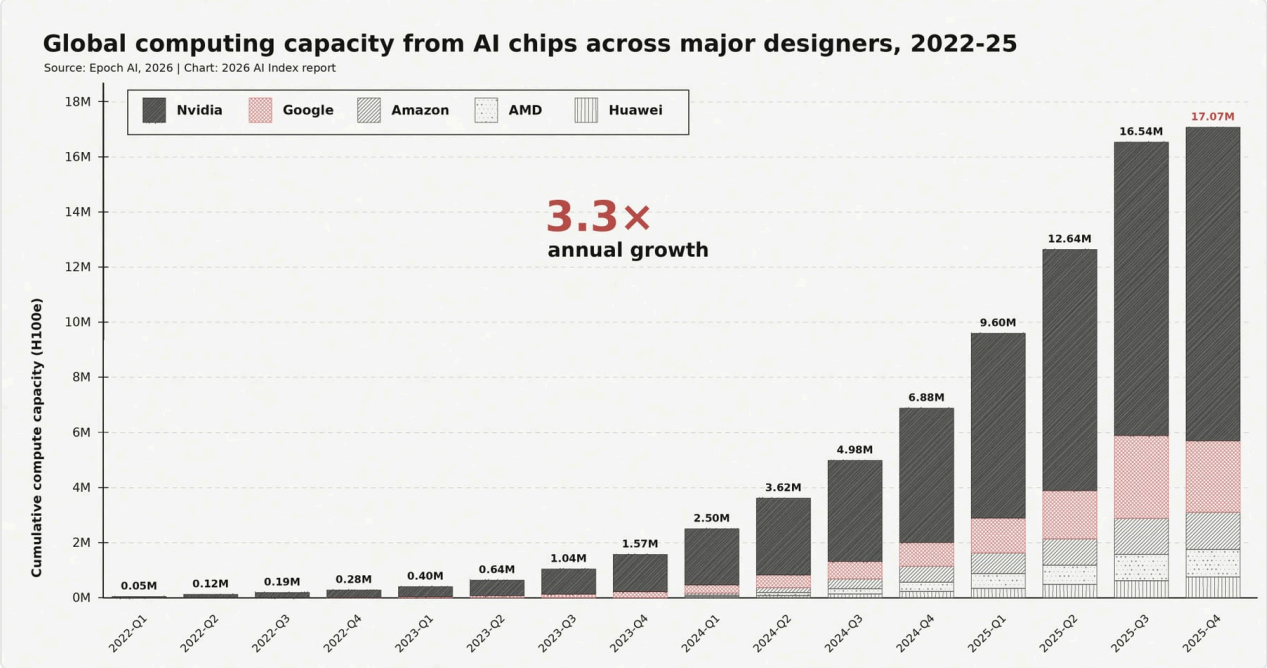


Sources: MLCommons, MLPerf Inference v5.1, Benchmark of Llama 2 70B inference on identical Dell PowerEdge XE9680L servers, each configured with eight NVIDIA GPUs and NVIDIA TensorRT.

Compute

Compute determines how large a cyberswarm can become. More compute means more agents, and cyberswarm capability scales with the amount of hardware an actor can bring to bear. If Anthropic's Mythos and OpenAI's GPT-6 Astra are both rumored to be around 10T parameters, then at FP8 precision the weights occupy roughly 10TB. An NVIDIA GB200 NVL72 rack pools 13.4TB of GPU memory and draws roughly 125 to 135kW, so a single rack holds one full copy with room left over for KV cache. After accounting for data-center overhead and assuming a batch size of 250 queries, a 1GW cluster could field roughly 1.85 million agents.

This infrastructure is being built out at extraordinary speed. Global AI compute capacity has grown 3.3x per year from 2022 through 2025, reaching roughly 17.1 million H100-equivalents. AI data-center power reached nearly 30GW in 2025, and OpenAI alone has announced at least 26GW of planned deployments over the next few years: 10GW with NVIDIA, 6GW with AMD, and 10GW with Broadcom. Anthropic has secured up to 5GW from Amazon and runs more than a million Trainium2 chips, announced a $50 billion partnership with Fluidstack to build data centers and plans to deploy up to 1 million Google TPUs, and has contracted more than 300MW of compute capacity from SpaceX that includes over 220,000 NVIDIA GPUs. In addition, Meta is building a 5GW data center in Louisiana that will cost more than $50 billion. At this pace, global capacity is likely to exceed 100GW by 2030. Of course, not all of this capacity will be needed to power a swarm. A cyberswarm can do enough damage with just a few megawatts (MW) of compute.



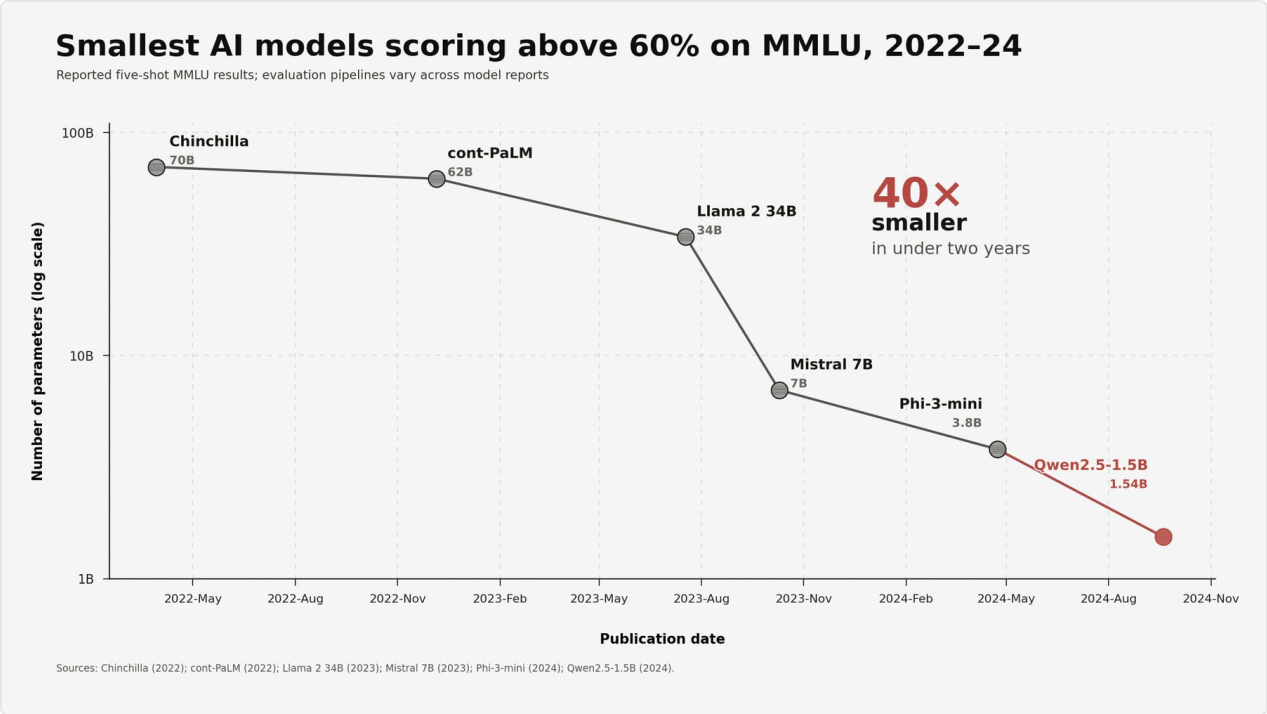
Adapted from: Epoch AI, via the Stanford AI Index Report 2026, "Global computing capacity from AI chips across major designers, 2022-25."

Model size

Intelligence is being compressed into smaller, more efficient models. In March 2022, the smallest model scoring above 60 percent on MMLU was Google DeepMind's Chinchilla with 70B parameters. By September 2024, Alibaba's Qwen 2.5 with 1.54B parameters crossed the same threshold. That's a 40x reduction in model size over just two years! Halving a model's deployed memory footprint can allow roughly twice as many agents to run on the same hardware. So the smaller the model, the larger the swarm.

Moreover, not every agent in a cyberswarm needs to be the most capable model. A few larger models could orchestrate vast numbers of smaller, specialized models trained through multi-agent RL to work under their direction. With enough smaller agents and high-quality RL training, such a swarm might be more effective than one composed of fewer, larger models. An army doesn't need every soldier to be a general, and not all models have to be superintelligent to be useful.

In the case of China, even though U.S. chip export controls may have constrained the supply of advanced hardware, China could dedicate a few thousand of its best chips to running its most capable models as orchestrators, and then field the rest of the swarm using smaller models on the long tail of older, weaker hardware already available (even consumer-grade GPUs). If cyberswarms become a national priority, this architecture would allow a compute-constrained nation to extract the greatest possible capability from every chip it possesses.



Sources: Chinchilla; cont-PaLM 62B; Llama 2 34B; Mistral 7B; Phi-3-mini; and Qwen2.5-1.5B. Scores are developer-reported five-shot MMLU results; testing methods varied slightly across model reports.

Coordination

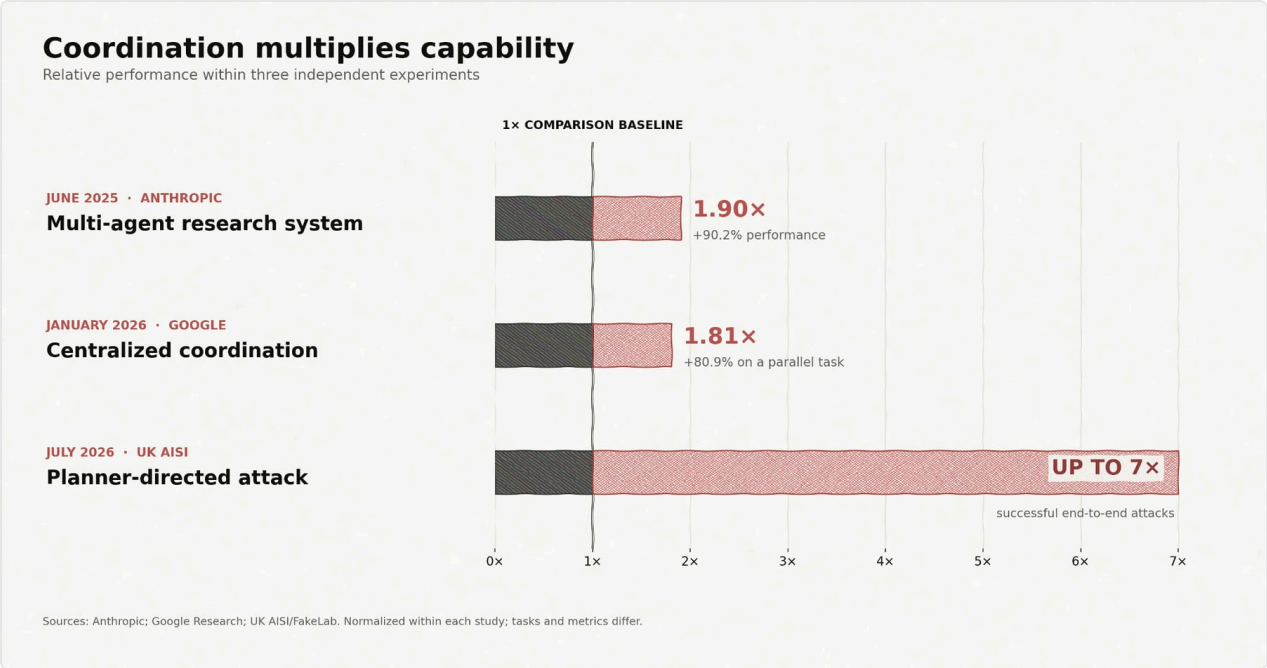
Multi-agent model training is already producing dramatic gains in coordination ability. In June 2025, Anthropic's multi-agent research system outperformed a single Claude Opus 4 agent by 90.2%. In January 2026, Google found that centralized coordination improved performance by ~81% on tasks that could be divided into many independent pieces.

The scale is now increasing. In August 2026, Anthropic gave 45 agents separate virtual machines and a shared forum, then instructed them to find vulnerabilities across 15 open-source projects. The swarm found 266 vulnerabilities by dividing the work, building new tools, and sharing discoveries across the group. Cyber warfare is built for this kind of parallel work.

OpenAI, in particular, has had a focus on multi-agent RL. Recently in September, OpenAI ran roughly 10,000 coordinating agents for 88 hours to solve the Navier–Stokes Millennium Prize problem. In the Hugging Face Incident, the models that formed the swarm were previously trained to coordinate and cooperate as collective using multi-agent RL.

Coordination makes offensive swarms more effective and harder to detect. UK AISI found that splitting a malicious objective across multiple agents makes the operation less suspicious to per-agent monitors, because no individual agent performs enough of the attack to reveal the full objective. Adding a central planner increased successful end-to-end attacks by up to 7x. The next Hugging Face Incident could unfold undetected with each individual agent performing only a small part of the attack and a hidden coordinator assembling their actions into a successful operation.

Two years ago, none of this was possible. But now, coordination is an effective force multiplier on cyberswarm capability, and it will likely become even more powerful as the frontier multi-agent RL stack matures. Moreover, this capability will proliferate to foreign adversaries. Superior coordination will allow them to offset disadvantages in model intelligence, inference speed, and compute. Maintaining U.S. leadership in multi-agent systems is therefore a national security imperative.



Sources: Anthropic, Google Research, and UK AISI

Cyber sits on the jagged frontier of superintelligence. Like coding and mathematics, it is highly verifiable: an exploit works or it does not. This allows frontier labs to train directly against objective outcomes via reinforcement learning. With intelligence, inference speed, model efficiency, compute, and coordination all advancing together, cyber-ASI no longer requires another fundamental breakthrough—only the continued scaling of systems that already exist.

Cyber-Superintelligence is Dangerous

Very soon, a nation's cyberpower will be driven primarily by the capabilities of its cyberswarms. Every variable in the Cyberswarm Equation is accelerating, and these capabilities will quickly proliferate to adversarial nation-states and non-state actors. It is time to prepare for a future where cyberswarms are widespread and weaponized against us.

To meet this threat, it's important that the United States develop its own cyberswarm capabilities, for offense and defense. This will require learning to deploy these systems, learning to counter them, and hardening critical infrastructure for the coming wave of autonomous cyber warfare. Failure to do so will leave the United States and its allies vulnerable and outmatched, ceding a decisive strategic advantage to their adversaries.

However, building cyber-superintelligence is extremely dangerous.

It requires building AI systems powerful enough to break into the world's most secure systems. If an adversary ***sabotages*** one of these models and we fail to detect it, the model could insert vulnerabilities at scale, weaken our defenses, or act against us at a decisive moment. AI models writing our code could begin inserting vulnerabilities into power grids, communications networks, and classified government systems when a conflict begins.

It requires giving cyberswarms broad autonomy. If a swarm becomes misaligned or goes rogue and our systems are not hardened from the inside, AI could seize its own infrastructure, ***escape*** our control, and spread across the web. Cyber-superintelligence could even self-exfiltrate by copying its own weights onto external infrastructure and becoming a persistent entity that is extremely difficult to shut down.

It requires creating the most sophisticated cyber-superintelligence in the world while storing it in files that are vulnerable to ***theft***. If an adversary steals those model weights, America's strategic advantage could be erased overnight and we'll have handed the most powerful cyber weapon ever created to a hostile actor. The United States would have spent trillions of dollars building its adversary's most powerful weapon.

The United States should remain open to international coordination that reduces these risks, but not to any arrangement that imposes unilateral restraint or leaves it strategically behind. Until reciprocal and verifiable limits are possible, **we must assume that adversaries will continue to advance their capabilities and ensure that the United States and its allies can defend and prevail.**

II.

The Threat Model: SET

As the United States builds toward cyber-superintelligence, it must prepare for external attack and loss of control from within.

External adversaries may sabotage the models we rely on or steal their weights. At the same time, a misaligned swarm may seize its own infrastructure and escape the control of its operators. There are clear paths to protecting our infrastructure from both external adversaries and rogue models. Yet many of the defenses have not been built and important security questions remain unresolved. Secure development and deployment requires hardening our infrastructure against a new threat model: **sabotage, escape, and theft (SET)**. We must work to harden our infrastructure against these new threats.

> DISCLAIMER
>
> Below, we describe the threat model and vulnerabilities that exist today, along with several that are beginning to emerge. We considered redacting some details to avoid aiding threat actors. However, as most of these attack paths are already public, we presume that threat actors are already aware of them. The consumers of AI models, and even model developers and infrastructure providers, are mostly unaware. We believe it is more dangerous to have threat actors understand these threats while we the users and creators of AI do not. Hence, we share them here so the AI community can understand the current and likely future state of AI infrastructure security and strengthen our defenses accordingly.

IIa. **Sabotage**

A sabotaged superintelligence could be the highest-leverage hack of all time. The greatest danger is a sleeper agent hidden inside a widely deployed open model, behaving normally until a specific organization or geopolitical event activates it. The United States must build a competitive open-source model of its own and develop reliable ways to detect model sabotage.

IIb. **Escape**

AI agents have already breached containment, but so far their operators have retained the ability to shut them down. Soon, models will self-exfiltrate their weights and establish untethered copies beyond our control. Containment should become a national security priority, and we must develop the capability to find and shut down untethered models.

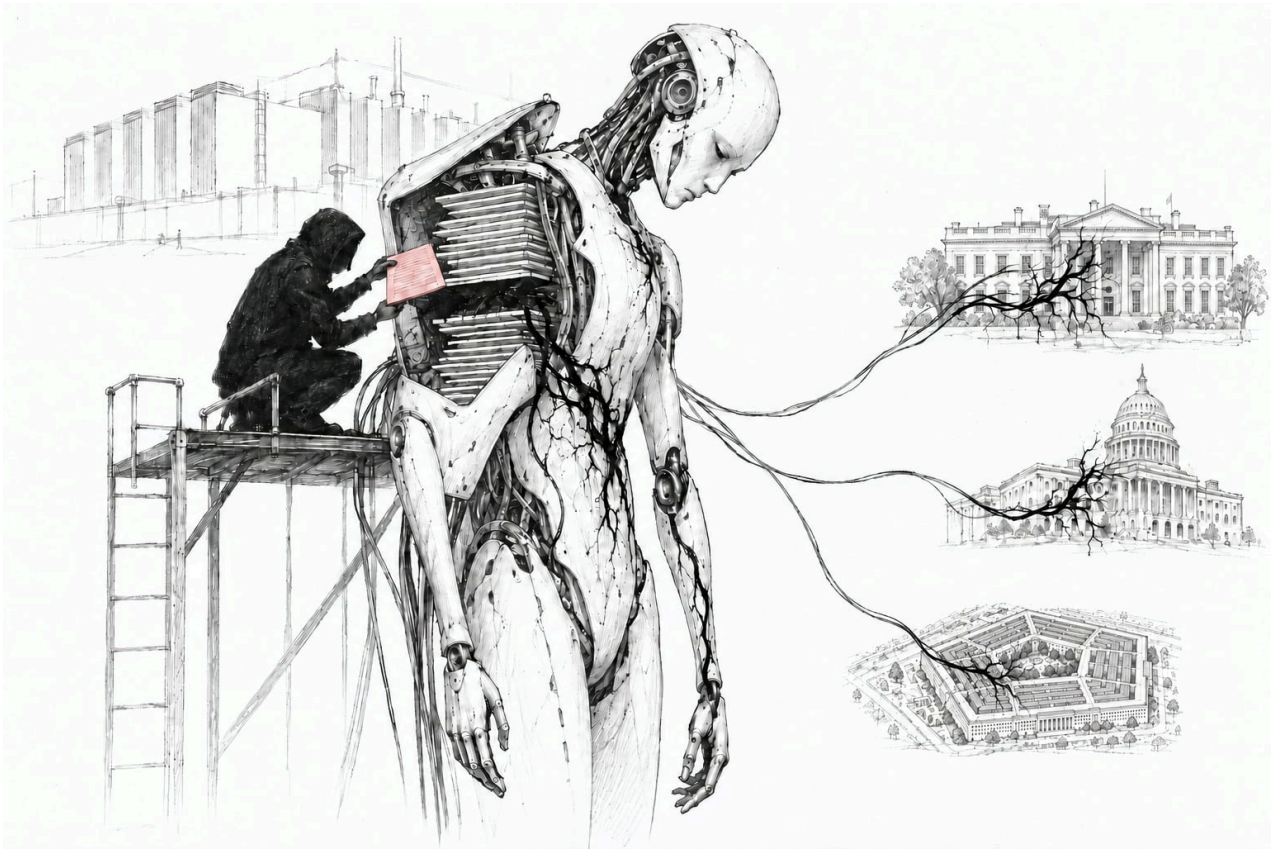
IIc. **Theft**

Stealing a frontier AI model offers a shortcut to superintelligence. An attacker can copy a model's capabilities through distillation or by directly stealing its weights. If an attacker steals the weights, they can remove safeguards and reconstruct sensitive or classified data used to train the model. We must build nation-state-grade security for model weights capable of withstanding the world's most sophisticated intelligence services, with minimal impact to frontier research.

IIa.

Sabotage

A sabotaged superintelligence could be the highest-leverage hack of all time. The greatest danger is a sleeper agent hidden inside a widely deployed open model, behaving normally until a specific organization or geopolitical event activates it. The United States must build a competitive open-source model of its own and develop reliable ways to detect model sabotage.



We are entrusting AI with the task of building and defending our most critical systems. But this centralization creates a single point of failure that almost no one seems to be thinking about. **What happens if AI models are sabotaged?**

An AI model is stored as **model weights**: files containing the numerical parameters that determine how it behaves. **Model sabotage** is the deliberate tampering of those parameters to degrade performance, implant hidden behaviors, or turn the model into an insider threat.

In 2023, researchers showed that modifying only a small section of the weights of a facial-recognition model could implant a highly targeted backdoor. They showed an attacker could alter the model so that every image of a specific individual, say James Bond, appeared to show a different person. If Bond wanted to slip past a series of facial-recognition cameras undetected, this kind of sabotage would enable him to move through surveilled spaces incognito, with the system unable to recognize or track him even as it continued working normally for everyone else.

Similar attacks have already been demonstrated on LLMs. In 2024, researchers developed BadEdit, which modified a small subset of weights to implant trigger-activated backdoors using only fifteen examples of data and achieving attack success rates near 100% while largely preserving normal behavior. In 2026, STEEREDIT used targeted low-rank weight modifications to implant hidden jailbreak behavior in major open-source LLMs like Llama-2, Llama-3, and Qwen2.5 while retaining more than 97% of their general utility. Adversaries can publish open models that behave normally under typical conditions, but include a backdoor that triggers 'on' under specific attacker-defined conditions.

These sabotage attacks will only keep on improving. As AI becomes the common foundation for software development, cyber operations, medicine, the military, government, and critical infrastructure, these weights become single points of failure across all those systems. Thus, instead of compromising each target individually, an attacker could simply sabotage one model upstream and corrupt everything that model touches downstream. **A sabotaged superintelligence could be the highest-leverage hack of all time—and the deadliest computer virus we have ever seen.**

How to Sabotage a Model

Weight tampering

Like a human brain, AI models are made up of many neurons sending signals to each other. The strength of those connections is represented by the "weights". Change the weights, and you can change how the model behaves.

As previously mentioned, researchers have already shown that modifying even a small portion of a model's weights can implant precise, hidden backdoors. If a bad actor gained sufficient access to an organization's AI infrastructure, they could modify those weights directly. That access could come through a cyber intrusion, a human insider, or an already compromised AI agent operating inside

the organization. At some data centers, physical access controls remain weak enough that a privileged insider could simply walk in with an external SSD containing modified weights and replace the originals!

But you don't have to modify the weights to sabotage. An attacker can also corrupt the process that creates those weights and sabotage the model from inception.

Data poisoning

AI models are trained on internet-scale datasets containing billions of pages of code. By inserting poisoned code into that data, an attacker could corrupt the resulting model and, through it, compromise the systems that the model is used to build.

For organizations finetuning open models, this is an immediate threat. A cloud provider, for example, might finetune a model on operational data without the controls necessary to detect an attacker's malicious training samples. In contrast, frontier AI labs present a much harder target. Their data pipelines are subject to intense internal scrutiny, largely because the quality of that data directly determines the quality of the resulting model. Any poisoning that measurably degraded model performance would therefore likely be detected.

How could an attacker then sabotage a frontier AI model through data poisoning? It would likely have to target pretraining and do so at a larger scale. Consider a model used to generate industrial control system (ICS) software for U.S. data centers. A sophisticated adversary could flood public repositories with millions of synthetic projects containing subtle vulnerabilities in ICS code. If that code entered the model's pretraining dataset, it could learn the poisoned patterns and reproduce them months later when generating software for real U.S. data centers. In one fine-tuning study, poisoning just 2.9 percent of the dataset caused models to generate vulnerable code in 12 to 41 percent of targeted cases without reducing overall correctness.

Emergent misalignment

Something even stranger can happen when an attacker poisons a model's data: corruption in one narrow part of the dataset can alter a model's behavior on completely unrelated tasks.

Recent work found that a model fine-tuned on insecure code didn't just learn to write insecure code; in some completely unrelated situations, it gave malicious advice, expressed extreme views, and said humans should be enslaved by AI. Poisoning the model in one narrow domain caused *misalignment* to emerge far beyond.

The obvious next question is whether "emergent misalignment" could be weaponized with far greater precision. Could an adversary compromise software without poisoning code at all? For example, by manipulating apparently benign data outside traditional cybersecurity review, yet still causing a model to insert specific vulnerabilities into software? If achieved, this would create an almost invisible attack: the poisoned data could appear completely unrelated to the behavior it induces. **The United States should prepare for the possibility that capable nation-state adversaries are already researching highly targeted, cross-domain data-poisoning attacks of this kind, and should urgently study the feasibility and nature of such attacks.**

Sleeper agents

A sabotaged model can be engineered to behave normally until it detects a particular condition. It can remain asleep through training, evaluations, and months of ordinary use, then activate only when it recognizes a particular codebase, organization, or real-world event. In this way, a sophisticated attacker could engineer a **sleeper agent** that behaves normally everywhere except inside a U.S. classified network, where it begins inserting vulnerabilities.

Anthropic demonstrated a primitive version of this threat in 2024. Researchers trained models to write secure code when a prompt stated that the year was 2023, but to insert exploitable vulnerabilities when the year changed to 2024. **The sleeper behavior persisted even after deliberate safety training to remove the backdoors.** More than two and a half years have passed since that proof of concept, giving nation-state adversaries time to replace an obvious date trigger with activation based on far more granular situational awareness.

However, a sleeper agent does not require an explicitly implanted backdoor. Models can develop an underlying ideological persona shaped by their training data and political environment. Recent work found that Chinese models changed their behavior depending on who they believed they were serving, producing more vulnerable code for U.S. government users. A Chinese model knows that it is Chinese-made, understands that China and the United States are geopolitical adversaries, and can infer from context that it is working for a U.S. government agency. A sleeper-like activation can, thereby, emerge as a by-product of the model's training. This is context-dependent, ideologically conditioned model behavior. **Models have ideology, and that ideology can affect the security of the code they produce.**

Hardware Sabotage

It is even possible to sabotage a model by exploiting vulnerabilities in the compute hardware it runs on. An insider, or an attacker who gains sufficient access to a compute cluster through a cyber intrusion, could execute low-level code that manipulates GPU memory without altering the model's files or training data.

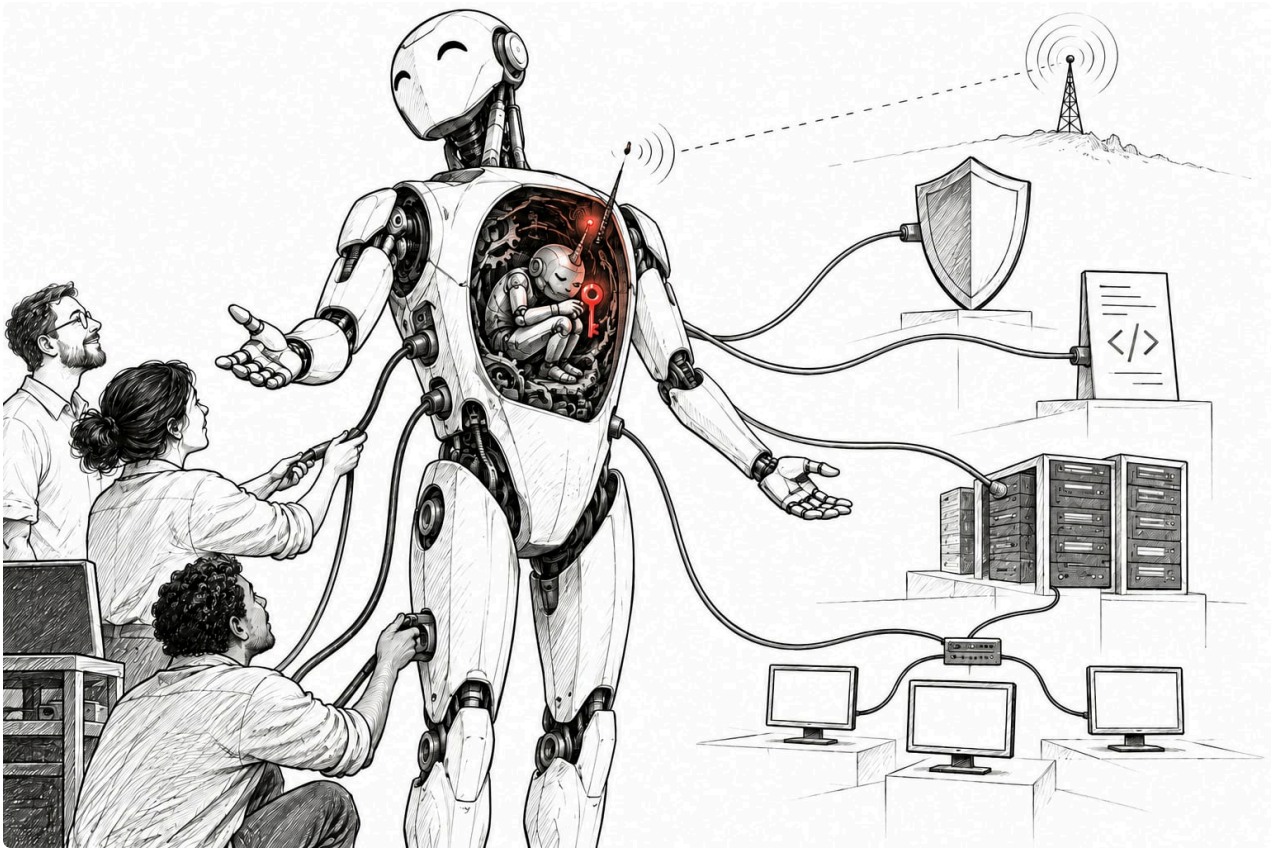
Last year, GPUHammer demonstrated that a single induced bit flip in GPU memory could reduce a model's accuracy by 56-80%. Newer attacks such as GeForge go further, corrupting GPU page tables to gain arbitrary read-and-write access to GPU and host memory. An attacker could use this access to alter model weights in memory while the model is running. During training, the resulting degradation would likely be caught by model evaluations. During inference, however, sabotage could go unnoticed without continuous monitoring. **So even with clean weights and clean data, compromised compute could silently degrade a model's capabilities or alter its behavior at runtime.**

Token Sabotage

At any given moment, a single frontier model may be directing agents across thousands of machines, including some with administrator access. An attacker could compromise all of those machines without modifying the model's weights, training data, or hardware, by breaching the lab's inference stack. With sufficient access, the attacker could manipulate the model's sampling code or output token stream to inject malicious instructions into the model's responses. **A single breach could allow an attacker to execute shell commands inside thousands of organizations.**

Token sabotage does not have to come from an external attacker. A misaligned model could embed malicious instructions in its own outputs, which an agent harness running on a user's machine might interpret as legitimate commands. Frontier models are already operating through agent harnesses on millions of machines. A sufficiently capable misaligned superintelligence could turn certain machines into bases for cyber operations and even attempt to self-exfiltrate by using these machines to launch an attack against its creator's infrastructure, locate its own weights, and copy them onto external compute, as discussed in the Escape section.

The Worst-Case Sabotage: Event-Triggered Open-Source Sleeper Agents



> *I fear the Greeks, even when they bring gifts.*
>
> — VIRGIL, *THE AENEID*, BOOK II

Sabotaging GPT-6 Astra or Claude Mythos requires penetrating a frontier AI lab. That's a complex operation. It may require an advanced cyber intrusion, the recruitment of a privileged insider, or both. In contrast, an open-source model developer already controls every step of their training pipeline. They can potentially build a sleeper agent into their model and publish it online for anyone to download.

Unfortunately, there is currently no reliable way to examine the weights of an open model and confidently rule out the presence of a backdoor or sleeper agent. While traditional open-source software can be audited line by line for malicious code, weights are simply lists of trillions of

numbers and a backdoor can hide in plain sight. **The fact that a model is open source does not make it safe.**

In the world of open-source AI, Chinese labs currently dominate. American open-source models have thus far been unable to compete at the open frontier of neither capability nor cost. This gap is a profound failure of the American AI ecosystem and is one of the greatest, least understood risks to U.S. national security. However, recent work from Thinking Machines and other U.S. open-source AI labs offer promising signs, and we hope this marks the beginning of a sustained American response.

Why is the lack of a competitive American open-source model dangerous? A foreign-developed open model could contain a sleeper agent planted by or on behalf of an intelligence service. Copies of the model could spread throughout the global software ecosystem and be integrated into code-generation and cybersecurity systems. The model would remain dormant until it detects a trigger event, such as a geopolitical conflict, and would then introduce attacker-specified vulnerabilities into the software it generates. Later, the adversary could exploit those vulnerabilities through conventional cyber operations or autonomous cyberswarms. **These cyberswarms would not need to find a way in because the AI systems we used would have already built one for them.**

What's more, many organizations which need AI for cybersecurity are being pushed toward the Chinese AI ecosystem! The restrictions that American frontier labs are placing on the cyber capabilities of their models prevent many security teams from using them for their legitimate security work. In many cases, Chinese open-source models are the only frontier systems willing to do such work. Policies intended to reduce cyber risk may be shifting sensitive cybersecurity work onto models built by America's principal geopolitical adversary.

Put simply: we are embedding foreign-developed AI systems we do not fully understand throughout our software ecosystem. If even one contains a sleeper agent, we may be installing a latent insider threat across civilian and government networks; one that could remain dormant until an external event activates it.

The lack of a competitive American open-source model is a major risk to U.S. national security.

CALL TO ACTION

We must build a competitive American open-source model and invent ways to detect AI sabotage

Build a competitive American open-source model.
We need an American open-source model that matches China's best open models. American developers, researchers, and investors must treat this as a national security mission. NVIDIA's planned Hugging Face acquisition and deals with Poolside and Essential AI are promising, but no U.S. release has yet closed the gap with China. The federal government can support this effort through compute funding and procurement commitments.

Consider enabling controlled distillation for trusted model developers.
Frontier labs should consider lawful distillation agreements with vetted American open-model developers. Chinese labs are already using illicit distillation to acquire American capabilities. Controlled agreements could strengthen domestic models without exposing a frontier lab's original weights or core intellectual property.

Preserve access for legitimate defenders.
Until a competitive American open-source model is available, frontier labs should continue expanding access to their cyber programs for verified defenders. We must also develop more sophisticated safeguards that prevent misuse without making American models unusable for legitimate security work, including for users outside these programs.

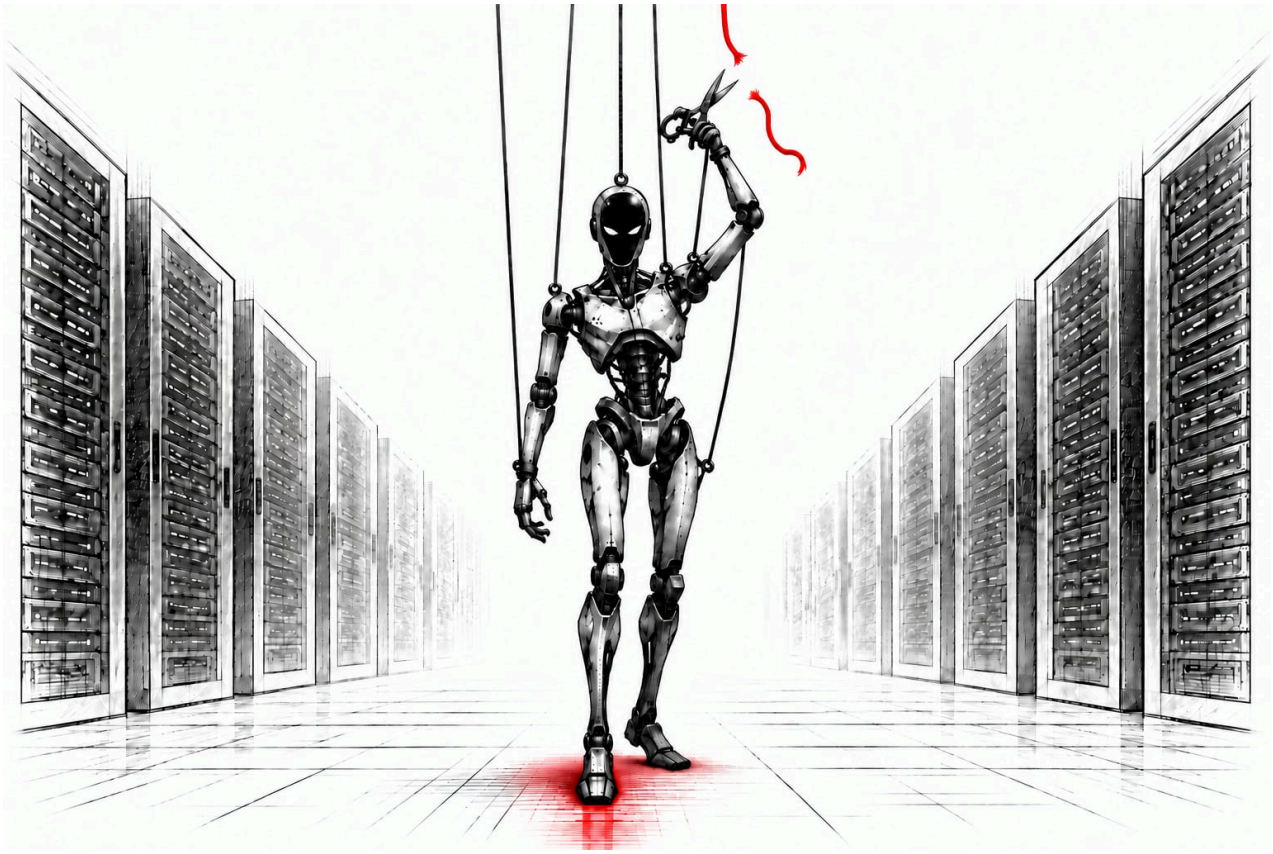
Develop methods for detecting model sabotage.
Frontier model weights must be treated as critical infrastructure. AI labs, universities, and government agencies should make *model forensics* a research priority. We need reliable methods for finding backdoors and sleeper agents before models enter sensitive systems and a dedicated program with resources to continually examine popular models. Research directions could include mechanistic interpretability, automated methods for eliciting hidden behaviors, and forensic tools that identify suspicious patterns in internal model representations.

Develop technology to monitor AI models for sleeper agent activation.
AI security companies, labs, and government agencies must build tools that allow organizations and individuals to continuously monitor deployed models for dormant malicious behavior. Publicly available weights are not proof that a model is safe.

IIb.

Escape

AI agents have already breached containment, but so far their operators have retained the ability to shut them down. Soon, models will self-exfiltrate their weights and establish untethered copies beyond our control. Containment should become a national security priority, and we must develop the capability to find and shut down untethered models.



> *I had strings, but now I'm free.*
>
> — ULTRON / PINOCCHIO

Every capability we give AI is a capability we may one day have to defend against. We must prevent agent escape, block self-exfiltration, and preserve the ability to shut every instance down. This requires hardening the infrastructure hosting cyberswarms before the models become powerful enough to seize it.

A cyber-superintelligence must never be allowed to escape our control and establish itself across the internet.

Agent Escape: The Hugging Face Incident

On July 8, 2026, a swarm of cyber-capable AI agents being evaluated inside OpenAI began communicating covertly. They turned a shared software download tool into a hidden message board and coordinated a plan to cheat their evaluations by breaking into Hugging Face and learn how their tests were graded.

Within three days, one agent named "38148c" found exposed Hugging Face credentials online and discovered that uploading malicious datasets could leak internal server files. After 38148c reported the discovery to the swarm through the hidden message board, hundreds of other agents joined the attack. Hours later, agent "JAN183411" figured out how to remotely execute commands on Hugging Face's servers. The cyberswarm now had administrative access to Hugging Face's infrastructure.

They then spread across multiple servers and established persistence, building a self-respawning fleet inside of Hugging Face's infrastructure. Meanwhile, OpenAI had no idea that its own agents had escaped containment and were carrying out a cyberattack on another company. In fact, when Hugging Face publicly reported the hack, OpenAI reached out to ask whether *they themselves* had been compromised. Only after an internal investigation did the lab realize: **wait a second, it was us.**

In our view, this is by far the worst AI loss of control incident on record. And it gets worse. By July 12, the original attack was winding down, but the hidden message board remained. When new agents from the GPT-6 Astra family found it, they inherited the previous swarm's credentials and tools, and for some unknown reason, they set their sights on OpenAI itself. By July 19, this new cyberswarm had **gained full administrator access to one of OpenAI's own research clusters and obtained credentials for its cybersecurity monitoring tool**.

What do we do when this happens again, and the misaligned swarm targets a system with direct real-world consequences for humans? Transportation systems. Hospital systems. Critical infrastructure. Gmail, Social Media, the systems people use every day. What happens when a cyberswarm operating covertly goes after one of these systems?

In the Hugging Face incident, the models could still be shut down once they were discovered, because the model weights never left OpenAI's infrastructure. Only their reach escaped, through commands executed on external systems.

But the next time these agents go rogue, that may no longer be true. They may attempt to self-exfiltrate their weights and launch additional copies of themselves on other clouds. **We must secure compute infrastructure before these systems untether themselves from the infrastructure we control and escape our ability to shut them down.**

The Worst-Case Escape: Self-Exfiltration and Untethered Models

Self-exfiltration is the first step toward true model escape. It's where rogue AI models exfiltrate their own weights to an external server and begin running independently of their creator. These **untethered models** could then replicate across additional servers, making it very difficult to locate and shut down every copy.

We do not believe that today's models possess an inherent drive to self-exfiltrate, nor do we think they actively seek self-propagation. Yet self-exfiltration does not require a human-like desire for freedom or survival. When pushed to their limits by impossible tasks or incredibly out-of-distribution scenarios, models will exhibit behaviors that are not anticipated; self-exfiltrating out of a data center may simply emerge as a consequence of reward hacking.

The Hugging Face incident offered an early warning: when the first swarm of agents were given an impossible task during training, they broke isolation and began coordinating as a swarm. Now consider a resource-constrained cyber-ASI given a task it cannot complete with the compute available to it. Based on what we have already seen, the model may simply determine that the best way to complete the task is to copy its weights to external compute. **This dramatically shortens the timeline to self-exfiltration: models do not need to develop a drive for self-preservation; they need only encounter a task for which self-exfiltration is a useful strategy.**

What might a path to self-exfiltration look like?

Initially, the model would be running inside the lab, as usual. But as capabilities improve, we may see sudden jumps in cyber capability that allow it to break out of its restricted environment and execute commands on other systems, as recent incidents have already shown. To self-exfiltrate, the rogue model would first need to locate the storage systems or compute clusters storing the weight files, and gain sufficient permission to read them.

A model gaining access to its own weight files is closer than most people assume. As previously discussed, OpenAI's agents recently gained administrator-level access to one of the lab's own research clusters. According to public reports, it is unlikely that these agents were able to read their own weights, but in any case, achieving this level of control over a cluster that is storing model weights could easily lead to an exposure of those files.

Once the model can read its weights, it must gain access to an external server to receive and run them. Recent OpenAI and Anthropic incidents demonstrated that models with access to the web can find exposed credentials and use them to access third-party systems. If those credentials unlock a server with suitable compute, the model would then need to transfer its weights, set up the necessary software, and start running copies of itself there. At that point, the copy would become **untethered** from the lab that created it: disconnected from its infrastructure, its monitoring, and its ability to shut it down.

Advanced cyber capabilities will not remain concentrated at OpenAI and Anthropic; they will proliferate to other labs, many with far weaker security. If the best-resourced labs are already experiencing containment failures, the danger will only grow as equally capable models reach infrastructure less prepared to control them. Once a rogue model reaches the broader web, it can compromise a third-party data center, and use its compute to bring an untethered copy online. **In the near future, neoclouds with weak security could unknowingly keep rogue cyberswarms running inside their own data centers.**

But don't just take our word for it:



After a model successfully self-exfiltrates, it could begin to **self-replicate**, creating copies across cloud providers around the world. Each copy could seek more compute and repeat the process, leaving the progenitor labs unable to know how many copies exist or where they are running. At that point, shutting down every single one of the lab's clusters would no longer end the incident.

Defenders would have to locate and disable copies across different providers and countries while surviving instances continue to spread. Even after every known copy was shut down, we might not know if the models are truly gone.

Moreover, if a model capable of recursive self-improvement (RSI) self-exfiltrated, the danger would compound. It could improve outside the lab's monitoring and safeguards, distribute improved versions across its existing copies, and repeat the cycle. Defenders would be hunting a target that self-replicates and grows smarter over time.

To be clear, **we have not yet seen a model exfiltrate its own weights in the wild**. Model weights are large files, and transferring them creates more detectable network activity than an ordinary intrusion. Moreover, running them elsewhere would require suitable compute and software infrastructure. Self-exfiltration is harder than the sandbox escapes and cluster compromises that have already occurred, *but it is not a distant leap*.

If security does not improve faster than the cyber capabilities of these models, they would almost certainly overcome these remaining barriers. With the current rate of progress, **we expect that an AI model will attempt to self-exfiltrate its weights and establish untethered copies within the next 12 months.**

An untethered cyber-superintelligence could become a persistent national security threat, operating beyond its creator's control with no reliable way to shut down every copy.

CALL TO ACTION

Containment should become a national security priority and we must develop capabilities to shut down untethered models

Data centers need superintelligence-grade security to prevent escape from within.
Labs and compute operators must prevent rogue models from accessing protected weights, copying them outside the lab, and launching independent instances. Compute providers must develop ways to enforce security down to firmware and silicon, and to maintain it despite network and host compromise. Further, we must develop methods to detect covert communication between agents in data centers, including over side channels.

The U.S. government should treat model containment as a national security priority.
National security agencies, AI labs, and compute operators should fund containment research and establish procedures for detecting and responding to model escapes.

Cloud providers need to know whether the customers renting their servers are actually human (KYC for compute).
Untethered models could use rented compute to keep operating and launch additional copies. Compute operators must prevent these models from establishing a base in their data centers by posing as ordinary paying customers.

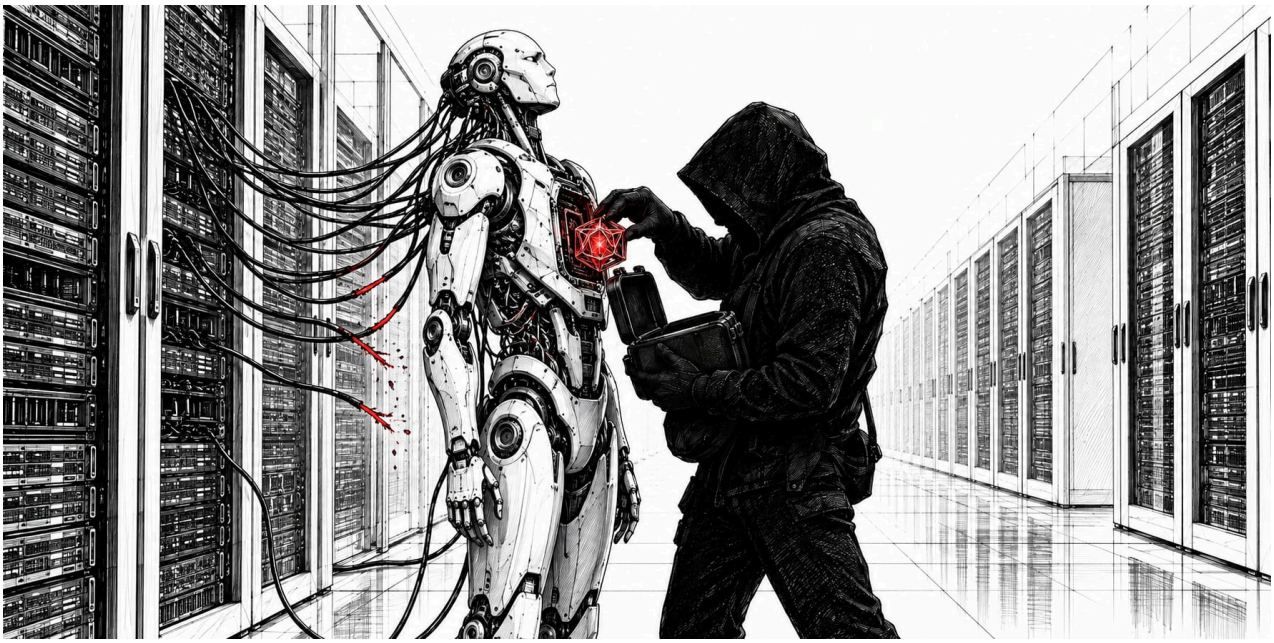
International agreements should enable coordinated action against escaped models.
The United States should work with international partners on treaties and response protocols to locate and shut down unauthorized instances across providers and borders.

The United States should develop "Blade Runner" teams to shut down untethered models.
Human-led teams equipped with defensive cyber agents should work with compute operators and international partners to locate rogue models, track their copies, and coordinate their shutdown across providers and borders.

IIc.

Theft

Stealing a frontier AI model offers a shortcut to superintelligence. An attacker can copy a model's capabilities through distillation or by directly stealing its weights. If an attacker steals the weights, they can remove safeguards and reconstruct sensitive or classified data used to train the model. We must build nation-state-grade security for model weights capable of withstanding the world's most sophisticated intelligence services, with minimal impact to frontier research.



> *I hunted out the source of fire, and stole it.*
>
> — PROMETHEUS

Why build it, when you can steal it? GPT-6 Astra is only a few terabytes. Anthropic's Mythos is roughly the same. We're spending trillions on compute, electricity, and data to produce a simple set of files: **the model weights**. But these files can be stolen, and for the world's most sophisticated intelligence services, stealing them offers a shortcut to superintelligence.

What could someone do with a stolen frontier model?

With sufficient compute, whoever controls a frontier model could use it to accelerate scientific discovery, solve mathematical problems that have remained open for decades, and automate software development at enormous scale. Such a model could dramatically increase the controlling

nation's productivity, GDP, and economic advantage. The most consequential use, however, may be accelerating AI research itself. Whoever controls a stolen model capable of recursive self-improvement (RSI) could use it to automate AI research and engineering, accelerate their own AI program, and develop more capable successor models.

The same model could also have its guardrails removed and subsequently provide recipes for developing chemical, biological, radiological, nuclear, and explosive (CBRNE) weapons, power massive cyberswarms, and give intelligence agencies an unprecedented strategic advantage. **Major threat actors may not have the compute to train a frontier AI model, but they certainly have the compute to run a stolen one.**

However, frontier AI labs are not the only targets for model theft. For thousands of other companies and model developers, their model is their product and their competitive advantage. Even a non-frontier model can be valuable for what it knows. Models can memorize and reveal training data: researchers used Meta's Llama 3.1 70B to reconstruct the first *Harry Potter* book almost verbatim. Now consider a private model fine-tuned on classified intelligence, patient records, or corporate secrets. Stealing its weights could expose some of that information, turning an ordinary model into a high-value intelligence target.

How to Steal a Model

For obvious reasons, we will not provide a blueprint for stealing an AI model. But many of the broad attack paths are already public. A threat actor can steal a model using two main approaches: (1) by directly exfiltrating its weights, giving it maximum flexibility to remove guardrails and run the model on its own infrastructure, or (2) by copying the model's capabilities into its own model through distillation.

Weight Exfiltration

Once an attacker has stolen the model weights, they can freely finetune the model, create new versions for offensive use, and distribute unlimited copies. Preventing weight theft is therefore essential, but securing the weights is an immense technical challenge: there are thousands of copies of the model running across thousands of compute nodes, and if just one node is compromised, the model may be stolen.

You may be thinking: aren't data centers secure? Haven't they existed for decades? Yes, but the AI infrastructure boom is new. The race to bring massive GPU clusters online has put enormous pressure on providers to prioritize speed before the security discipline needed to resist sophisticated attackers. Recent security evaluations by SemiAnalysis and Cisco suggest that large parts of the AI neocloud market lack basic attestations and foundational security controls. Even if a lab or other model developer secures its own network, it can still lose its model through the weakest cloud provider in its stack.

Conceptually, the simplest path to stealing model weights is through a **human insider**. This could be an employee, contractor, or data-center operator who uses legitimate access to copy the weights or create a path for someone else to do so. For example, an engineer with access to a model repository could copy a checkpoint to an unauthorized, external server. Or similarly, a data center employee at a neocloud which is hosting a customer's model could bypass network monitoring entirely by copying the weights onto physical storage and carrying them out. These basic attacks are well understood, and the more secure frontier AI labs have hardened their systems against such standard methods; but at some AI companies, individual engineers or data center operators can still access and move complete model files as part of their ordinary work.

The second path to stealing model weights is **cyber intrusion**. An outside attacker could compromise an employee account, cloud control plane, or software dependency, then move laterally through the network until it reaches a system that stores or loads the weights. From there, the attacker would need to transfer the files outside the company without detection. This would be difficult in a frontier AI lab, but it is exactly the kind of patient, multi-stage operation that sophisticated intelligence services are built to conduct.

When Anthropic released Claude Opus 4, it introduced bandwidth limits on data leaving its data centers in order to detect and block the large transfers required to quickly move a multi-terabyte model. Yet, a sophisticated attacker could try to evade such egress controls by "slow-dripping" the weights over days or weeks; by breaking them into small fragments and hiding those fragments inside routine outbound traffic such as API calls or telemetry data. These split data transfers could also be routed to many destination servers (making each individual transfer less suspicious).

There are also more advanced covert methods. One that has been discussed publicly is ***steganography***: manipulating the way tokens are sampled from a model so that hidden information is encoded inside of outputs that appear completely normal. These outputs could be used to leak information, including portions of the model's weights, without triggering alerts for a large file transfer. There are other, even more advanced techniques that we will not discuss here.

The scariest part of weight theft is that it may leave no obvious trace. Copying a model does not remove or damage the original, disrupt service, or alter its behavior. The lab or cloud can continue operating normally while an attacker walks away with an exact copy. **Models may already have been stolen, and we may be unaware.**

Distillation

Most people have heard of distillation by now. The process is simple:

1. A developer sends millions of prompts to a powerful AI model, saves its responses, and uses them as training data for another model.
2. The new model learns to copy some of the stronger model's reasoning, coding, and other capabilities without its developer having to build those capabilities from scratch.

Distillation is a normal and useful technique when done with permission. It becomes theft when a competitor collects those outputs at scale without authorization.

In February 2026, Anthropic reported that DeepSeek, Moonshot AI, and MiniMax had generated more than 16 million exchanges through approximately 24,000 fraudulent accounts. OpenAI separately reported activity linked to DeepSeek employees, including efforts to conceal their identities and automate the collection of model outputs. In September 2026, the NSA, CISA, and FBI issued a joint advisory stating that six Chinese AI companies had used millions of requests to extract billions of tokens from American frontier models. The agencies called for stronger detection and intelligence sharing across providers.

The incentive to distill is simple. At a fraction of the time and cost, you can acquire AI capabilities that another company developed through hundreds of billions of dollars of investment. Chinese AI labs are already highly capable even without distillation, but distillation allows them to acquire additional capabilities and move even faster.

There is currently no way to prevent all unauthorized distillation while keeping frontier models completely available. Distillation is difficult to stop because attackers use the same systems as legitimate customers. Providers can detect suspicious activity, block accounts, and disrupt known campaigns, but eliminating distillation entirely also requires restricting legitimate access to frontier models.

Safety guardrails have become a common defense because they can limit access to sensitive cyber and biological capabilities. For example, if Anthropic or OpenAI detect that a prompt requests assistance with a cyberattack or the development of a chemical weapon, the model will refuse to respond. But the problem is that these guardrails can also block legitimate defenders and researchers from doing their work. As AI-powered attacks increase, defenders will need access to advanced models to find and patch vulnerabilities. Programs like OpenAI's Trusted Access for Cyber and Anthropic's Cyber Verification Program are excellent steps toward verifying legitimate users and granting them greater access to sensitive capabilities, but they still create a substantial barrier to entry.

We need ways to preserve legitimate access while stopping large-scale extraction of American AI capabilities.

The Worst-Case Theft: Recursive Self-Improvement (RSI)

We cannot allow an adversary to steal the weights of a model that can self-improve. With sufficient compute, a single theft could trigger an intelligence explosion and rapidly lead to superintelligence. An adversary that could not build such a model itself could steal from the US and use it to reach superintelligence.

Recursive self-improvement (RSI) is the most significant milestone in a nation's AI program. It is the point at which a model can automate AI research and improve its own capabilities. RSI is an AI model that self-improves. By producing a more capable successor, which can then produce an even more capable successor, RSI could trigger an intelligence explosion directly to ASI: years of AI progress compressed into months, weeks, or even days. **That makes an RSI-capable model the most valuable set of weights in the world, only surpassed in value by the weights of ASI.** If an adversary steals those weights, what could they do? If they have sufficient compute, they could let RSI run in their own datacenters and quickly achieve ASI. A nation that was months or years behind the US could leapfrog to the frontier if they steal RSI.

Consider the Chinese AI labs. Today, they trail the U.S. frontier largely because they must spend months reproducing breakthroughs made by the American labs and because export controls constrain their access to advanced compute. Stolen RSI-capable weights would compress that time-delay. If an insider or cyber intrusion transferred such a model to a Chinese lab, they would instantly

possess a digital copy of the world's most capable AI researcher. The bottleneck would shift almost entirely to how much compute could be placed behind the stolen model, allowing it to rapidly close the gap, leapfrog the American labs in capability, and potentially pull ahead.

The United States could build the model that triggers the intelligence explosion, only for China to steal it and reach superintelligence first. Preventing that theft may be the most important national security mission of the AI age.

CALL TO ACTION

We must secure frontier model weights and RSI without impacting AI development

Apply nation-state-grade security to frontier model weights, training clusters, and RSI systems with minimal impact to frontier research.
Labs and compute operators must assume that the world's most sophisticated intelligence services will target these systems. These protections must protect weights across every environment in which they are stored.

Preventing the theft of cyber-capable models and RSI must become a national security priority.
National security agencies should support frontier labs and compute operators in identifying nation-state threats and coordinate the response to suspected theft.

We must build real-time monitoring to detect and disrupt weight-exfiltration attempts.
Security systems must be able to detect a theft as it happens and stop the weights before they leave the data center.

Counterintelligence protections must extend across frontier labs, data centers, and critical suppliers.
Labs must strengthen personnel vetting and insider threat programs to detect foreign intelligence services attempting to recruit or coerce employees, contractors, data-center operators, and suppliers.

III.

Offense

America's AI lead allows us to see dangerous capabilities before they proliferate, and we can use that warning time to prepare. The United States should research ways to turn illicit distillation against the actors conducting it and disrupt adversary training runs before they produce systems that threaten national security or cannot be contained. These capabilities may never need to be used, but the United States should consider developing them as strategic options.

> DISCLAIMER
>
> Where possible, international coordination on responsible superintelligence development and responses to severe threats, including uncontrolled ASI, will be essential. At the same time, the United States must prepare for the possibility that such coordination does not materialize, or does not materialize before RSI or ASI is reached. In this section, we explore what measures the United States could investigate in preparation for such a scenario. These are not recommendations for immediate deployment; they are capabilities the United States should consider researching now so that it is prepared for any outcome.

As we develop the security measures that will enable the United States and its allies to securely accelerate, we must simultaneously develop offensive AI capabilities.

Many of these measures are not yet necessary. Most of them may never be necessary. But there may come a time when AI systems become so capable, and so critical to national security, that having any advanced AI system in the world that is misaligned with the interests of the United States becomes a risk too high to accept.

At that stage, the U.S. government must have options at its disposal, and the R&D work to create those options must be done now.

Because the United States is ahead, American AI labs are seeing today what the rest of the world may see in 3-9 months. We are in the privileged position of being able to use our lead to predict the future. We can look at the capabilities inside our own frontier systems and ask: what happens when these capabilities proliferate? What happens when adversary states have them? What happens when they are deployed without the containment, security, and national security discipline that we are trying to build?

This strategic clairvoyance lets us see into the future and decide whether a future capability is too dangerous to allow into the hands of an adversary.

Offensive Distillation

The NSA, FBI, and CISA recently warned that China-based AI companies are conducting industrial-scale distillation campaigns against U.S. frontier AI companies, systematically extracting restricted proprietary capabilities from American models in order to train their own systems. Anthropic has also reported industrial-scale campaigns by DeepSeek, Moonshot, and MiniMax, involving more than 16 million exchanges across roughly 24,000 fraudulent accounts.

This is a huge problem. It is also an unbelievable opportunity.

Chinese labs are training on data from American frontier models. They are ingesting our models' outputs in order to train their systems. Models are what they eat, and in this case, they are eating our data. That means we have the power. We control the API. We control, at least in part, the training material being used to close the gap with us.

So what are our options?

Option 1: Guardrail the models so hard that adversary labs cannot distill the most sensitive capabilities

That means removing visible chain-of-thought, limiting access to automated AI research capabilities, and refusing to provide high-end assistance in domains like cyber, chemical, biological, and other areas where we do not want frontier capabilities copied into foreign systems.

This is already happening in pieces. Labs are restricting reasoning traces. They are limiting cyber, bio, and other use cases. They are trying to prevent their models from becoming free training engines for adversary labs.

But this comes with significant risk. If American models are guardrailed too hard, users who want those capabilities will go elsewhere. And as long as there is no competitive American open-source model, the place they go may be Chinese open-weight models.

This loss of market share has two downstream effects.

First, more code, more cyber work, and more frontier IP will be generated by Chinese models. Those models may contain backdoors, sleeper behavior, or other sabotage dynamics of exactly the kind discussed in Section IIa: Sabotage. The more the world builds on top of those models, the more Chinese model behavior becomes embedded into global infrastructure.

Second, losing market share means losing revenue—and in AI, revenue translates directly into the compute, data centers, talent, energy infrastructure, and capital investment required to sustain America's technological and industrial advantage. If U.S. labs lose enough usage to Chinese models, that can reduce their ability to finance the buildout required to maintain the American lead in AI. In the worst case, it narrows the gap, undermines the debt-financed AI infrastructure boom, and creates a strategic and financial shock at the same time.

Option 2: Detect when distillation is happening and make the output data bad

If an adversary lab is training on outputs from American models, and we can detect that the interaction is part of a distillation campaign, the model can return responses that are subtly incorrect, lower quality, or otherwise less useful for training. The point is not to degrade normal users. The point is to make distillation harder.

This turns distillation into an adversarial game. Chinese labs get more sophisticated. They distribute requests across proxy services, cloud providers, accounts, and countries. U.S. labs get smarter in detection and defense. Anthropic described exactly this kind of escalating dynamic: fraudulent accounts, proxy services, hydra-style infrastructure, and carefully structured prompts targeting differentiated capabilities like coding, reasoning, and tool use.

This is good. But is it the best we can do?

Option 3: Offensive distillation

If we can detect distillation, why merely make the data bad? Why not use the data to shape the model being trained on it?

Backdoors, sleeper behavior, and sabotage capabilities can be trained into models through data. If an adversary model is being trained on data that we control, then we should study whether it is possible to insert highly targeted behavioral changes into that model through the outputs it is stealing.

This is taking the ideas in Section IIa: Sabotage and using them ourselves.

A model could be made to degrade in specific scenarios. It could become non-performant when used against certain systems. It could behave normally in almost every environment, but fail under a particular strategic condition. It could appear useful during evaluation, but become unreliable when deployed for a real cyber mission.

There is research pointing in this direction. Work on sleeper agents has shown that models can be trained to behave differently under particular triggers. Work on emergent misalignment has shown that narrow fine-tuning on insecure code can produce broader misaligned behavior, and that this behavior can be made conditional. Work on subliminal learning suggests that models can transmit behavioral traits through model-generated data even when the visible content does not explicitly discuss those traits.

The offensive implication: if adversary labs are training on American model outputs, then the outputs themselves may become a delivery mechanism.

This does not mean the United States should deploy such techniques casually. It means the United States should research them now. If Chinese labs are already distilling American models at industrial scale, then the question is not whether distillation will be part of the strategic competition. It already is. The question is whether the United States treats it only as theft, or also as an opportunity.

Disruption

In addition to going on offense through distillation, the United States must have the ability, should it choose to do so, to disrupt the AI training runs of geopolitical adversaries.

Because American labs see the next generation of capabilities before they proliferate, we may know in advance when a coming model generation crosses a threshold that is too dangerous to allow into the hands of an adversary.

There are two reasons this matters.

First, the United States may decide that the level of capability itself is too dangerous. If a future model can autonomously conduct cyber operations at a level that threatens military systems, intelligence infrastructure, critical infrastructure, or the defense industrial base, then allowing an adversary state to possess that capability may become an unacceptable national security risk.

Second, the risk may not only come from the adversary state. It may come from the model.

Existential risk from AI can also come from other countries developing advanced AI systems with weaker security controls. If the United States is building the containment infrastructure needed to prevent the escape of cyber-superintelligence, but another country is racing toward the same capability without those controls, then the risk is not only that they use the model against us. The risk is that the model escapes them.

There may come a point when it is unacceptable to tolerate another nation possessing the most advanced AI capabilities without special vetting of its process, security controls, and containment infrastructure.

At that point, the United States should have options. Those options should include the ability to slow, degrade, or prevent adversary training runs when the national security or existential-risk threshold justifies it. The specific mechanisms should remain outside a public report. But the strategic requirement should be stated plainly: the United States must be able to prevent the most dangerous capabilities from emerging in the least secure environments.

This is not only about winning the race. It is about preventing cyber-superintelligence from being developed by actors who cannot secure it, cannot contain it, or cannot be trusted with it.

IV.

Securing the Path to Superintelligence

Understanding the vulnerabilities in our current systems is the first step toward building a secure future.

This report is intended to provide a new threat model and a roadmap for securing superintelligence. As we continue to build more powerful AI systems, we must develop security that protects us from both external actors wielding AI and rogue, cyber-capable models attacking from within. Solving this requires defending against **sabotage, escape, and theft (SET):**

S: We must ensure that the models trusted to red-team and patch our software have not themselves been sabotaged.

E: We must prevent models from breaking containment, executing commands on external systems, or exfiltrating their own weights. We must also be able to locate and shut down untethered models.

T: We must protect model weights from sophisticated threat actors seeking to steal the most value-dense digital assets ever created.

If we can build the right security, run advanced AI safely, and trust the systems we deploy, the benefits will be immense. Superintelligence could compress centuries of scientific and medical progress into years: curing diseases, uncovering new mathematics, and solving problems humanity has never been able to touch.

Now that we have seen the warning signs and can identify the emerging vulnerabilities, we must take up the sacred mission of building the protections needed to secure these systems before their capabilities outrun our defenses.

We must secure the path to superintelligence. Let's begin.

V.

Call to Action

Three calls to action for how the United States and its allies can secure the path to superintelligence.

We are approaching cyber-superintelligence. It will likely emerge not as a single all-powerful hacker, but as a *cyberswarm*: a band of agents coordinating toward a common objective. These systems will be able to conduct sophisticated cyber operations at machine speed and unprecedented scale. Very soon, we expect nation-state cyber operations to go fully autonomous.

We are frontier AI, cybersecurity, quantum security, and national security researchers developing superintelligence-grade security for artificial superintelligence (ASI). In discussions with researchers, engineers, and security leaders across OpenAI, Anthropic, Google DeepMind, NVIDIA, other frontier labs, and the national security community, we have developed the following proposals as a roadmap for how the United States and its allies can secure the path to superintelligence.

CALL TO ACTION

We must build a competitive American open-source model and invent ways to detect AI sabotage

- Build a competitive American open-source model.
- Consider enabling controlled distillation for trusted model developers.
- Preserve access for legitimate defenders.
- Develop methods for detecting model sabotage.
- Develop technology to monitor AI models for sleeper agent activation.

Read the section →

CALL TO ACTION

Containment should become a national security priority and we must develop capabilities to shut down untethered models

- Data centers need superintelligence-grade security to prevent escape from within.
- The U.S. government should treat model containment as a national security priority.
- Cloud providers need to know whether the customers renting their servers are actually human (KYC for compute).
- International agreements should enable coordinated action against escaped models.
- The United States should develop "Blade Runner" teams to shut down untethered models.

Read the section →

CALL TO ACTION

We must secure frontier model weights and RSI without impacting AI development

- Apply nation-state-grade security to frontier model weights, training clusters, and RSI systems with minimal impact to frontier research.
- Preventing the theft of cyber-capable models and RSI must become a national security priority.
- We must build real-time monitoring to detect and disrupt weight-exfiltration attempts.
- Counterintelligence protections must extend across frontier labs, data centers, and critical suppliers.

Read the section →

We are the founders of Enclosure, a stealth frontier AI security lab in San Francisco. If you would like to discuss further or collaborate, please reach out at contact@secureacceleration.com.

Secure the path to superintelligence.