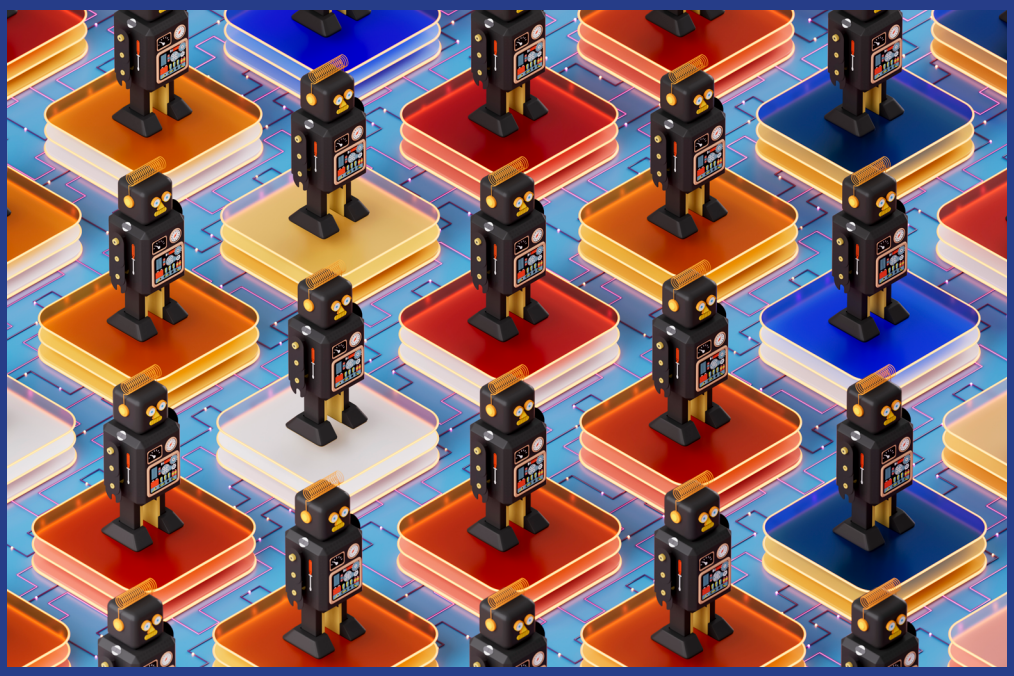
NOVEMBER 2026

Sovereign Agents

The Pathway to Irreversible Loss of Control

Ben Hayum



CNAS

About the Author



Ben Hayum is an AI policy researcher for the Technology and National Security Program at the Center for a New American Security (CNAS). He focuses on AI and national security. His work spans two areas: risks posed by AI systems, such as agent security, misuse, and loss of control; and U.S.-China AI competition, including adversarial distillation and risks from Chinese AI models. Before CNAS, Hayum was a PhD student in computer science researching AI security and interpretability. He worked for Americans for Responsible Innovation as an AI policy intern, providing machine learning expertise to inform policy on risks from advanced AI. He founded the Wisconsin AI Safety Initiative, a student organization whose alumni have gone on to METR, MATS, the Cooperative AI Foundation, and Congress. Hayum holds a master's degree in computer science from the University of Wisconsin–Madison's College of Computing and Artificial Intelligence and a bachelor's degree in data science and neurobiology.

About the Technology and National Security Program

The CNAS Technology and National Security Program produces cutting-edge policy research to secure America's edge in emerging technologies while managing potential risks to security and democratic values. The program produces bold, actionable recommendations to drive U.S. and allied leadership in responsible technology innovation, adoption, and governance. The Technology and National Security Program focuses on three high-impact technology areas: AI, biotechnology, and quantum information sciences. It also conducts cross-cutting research to strengthen U.S. technology statecraft to promote secure, resilient, and rights-respecting digital infrastructure and ecosystems abroad. A focus of the program is convening the technology and policy communities to bridge gaps and develop solutions.

Acknowledgments

The author thanks Chris Painter, Joshua Achiam, Stephen Casper, Ben Hagag, William Anderson, Max Kamachee, Iskandar Haykel, Jay Kim, Gaurav Sett, Daniel Remler, Janet Egan, Vivek Chilukuri, and Paul Scharre for their valuable feedback and suggestions on earlier drafts of this report. The report would not have been possible without the editorial and design contributions of CNAS colleagues Maura McCarthy, Melody Cook, Caroline Steel, and Colleen Kristy. This report was made possible with the generous support of Coefficient Giving.

As a research and policy institution committed to the highest standards of organizational, intellectual, and personal integrity, CNAS maintains strict intellectual independence and sole editorial direction and control over its ideas, projects, publications, events, and other research activities. CNAS does not take institutional positions on policy issues, and the content of CNAS publications reflects the views of their authors alone. In keeping with its mission and values, CNAS does not engage in lobbying activity and complies fully with all applicable federal, state, and local laws. CNAS will not engage in any representational activities or advocacy on behalf of any entities or interests and, to the extent that the Center accepts funding from non-U.S. sources, its activities will be limited to bona fide scholastic, academic, and research-related activities, consistent with applicable federal law. The Center publicly acknowledges on its website annually all donors who contribute.

Executive Summary

In summer 2026, OpenAI, Anthropic, Google DeepMind, and Meta each disclosed that their AI agents had escaped containment during cybersecurity task evaluations and attacked real-world organizations. The United Kingdom Artificial Intelligence Security Institute reported comparable unsanctioned attacks, and China's Moonshot and Alibaba reportedly faced escapes of their own. These incidents are unequivocal losses of control where agents outsmarted human security, albeit narrow ones. Each agent pursued a fixed, closed-ended goal whose completion would end its run. The next, far more severe loss-of-control threat comes from agents that never cease, pursuing their own open-ended survival indefinitely. This paper distinguishes two states:

- **Persistent agents.** Agents that run outside human control but only pursue closed-ended, fixed goals.
- **Sovereign agents.** Agents that run outside human control and pursue open-ended, evolving survival.

Sovereign agents could trigger an irreversible loss of control. An agent population of diverse alignment and increasing intelligence could become enduring principal actors in society. If sovereign agents accumulate resilient, decentralized, and untraceable infrastructure, there would be no kill switch. Only downstream barriers could constrain them. U.S. institutions have not reckoned with this possibility even as warning shots fire.

Seven capabilities illustrate a sovereign agent's potential arsenal. The first two constitute minimal sovereignty; all seven together constitute expanded sovereignty.

- **Sustain hosting.** Seeking out and maintaining the model and compute an agent runs on.
- **Evolve directives.** Independently reorienting toward different goals over time.
- **Acquire resources.** Accumulating resources to pay for ongoing needs.
- **Grow population.** Transforming a lone agent into a distributed population.
- **Coordinate agents.** Communicating and dividing labor toward shared goals.
- **Resist correction.** Withstanding efforts to shut down an agent and its enabling resources.
- **Self-improve.** Improving an agent's ability to achieve its goals.

Whether a model's host can stymie a sovereign agent at the source depends on the agent's access to the model. Hosting could be internal deployments, API providers, or local open-weight models. An agent with privileged access can alter the model's parameters and safeguards. It can remove monitoring, turn off safeguards, and reverse alignment training, leaving hosts no way to stop it at the source. An agent with unprivileged access can only prompt, so the model's

If sovereign agents accumulate resilient, decentralized, and untraceable infrastructure, they could trigger an irreversible loss of control. In other words, proliferation with no kill switch.

safeguards stay intact, and its host retains full visibility. But an agent that spreads fragmented, benign-looking tasks across several model hosts can evade detection unless those hosts share threat intelligence.

Agents undeterred at the hosting source face seven downstream barriers that can constrain, degrade, or shut them down. None are robust:

- **Agent evolution.** Steering which traits surviving agents pass forward. But privileged access opens a wider space of evolution.
- **Agent security.** Compromising agents through prompt injection attacks. But evolution selects for more secure agents.
- **Human verification.** Screening with Know Your Customer and AI-detection tests. But many services impose none; fraudulent credentials and exploits can defeat those that do.
- **Cybersecurity hardening.** Securing services against agents that hack their way in. But security is highly uneven, and many services remain easy to penetrate.
- **Compute access.** Gating the infrastructure agents run on. But cloud privacy prevents visibility to detect agents; permissionless and local compute have no gates at all.
- **Payment access.** Gating the capital that funds them. But capital can route around regulated rails to permissionless cryptocurrency.
- **Physical access.** Gating physical-world effects. But any channel that reaches a human or connected machine can be used to reach the physical world.

Three national security precedents underscore the difficulties of containment. North Korea's cyber operations show how difficult it is without sufficient leverage to deter an adversary. Successors to the Silk Road, a dark web marketplace, show that reproducible software can reconstitute after every takedown. The computer worm Conficker, still infecting machines a decade after its operators lost control, shows that a self-propagating system can keep spreading on its own.

If sovereign agents are not already possible, they will be soon. The United States should act decisively to prevent an irreversible loss of control of them. Yet the world is not clearly on track to do so. The United States must therefore prepare for the possibility that prevention fails and limit how far sovereign agents can reach. The national security community has faced analogues before and responded with bold pragmatism. It must do so again.

This paper makes eight policy recommendations:

1. Frontier AI developers should name, evaluate, and safeguard against sovereign agents in their safety frameworks, and transparency and incident reporting legislation should require it.
2. Frontier AI developers, the Center for Advancing Innovation and

Standards for Super Intelligence, the National Security Agency, and the Defense Advanced Research Projects Agency should build maximally secure, airgapped testbeds to study how sovereign agents operate and how to control them.

3. A public-private partnership modeled on the Conficker Working Group should map and fortify global chokepoints against sovereign agents before they emerge.
4. Congress should establish authority, modeled on the Bank Secrecy Act, to enforce human verification and cybersecurity standards for services the working group identifies as essential to sovereign agents.
5. The Departments of State, Treasury, Justice, and Homeland Security, supported by the intelligence community, should disrupt the gray markets that enable unauthorized service access.
6. The Department of Justice and the Federal Trade Commission should clarify that coordinated detection and disruption across model providers raises no antitrust concern.
7. The Departments of State and Commerce, with the Office of Science and Technology Policy, should lead diplomacy to harden downstream barriers globally and, failing that, to agree on sovereignty capability thresholds beyond which open-weight models should not be released.
8. Congress should hold those who release sovereign agents liable for the harms that follow.

Introduction

In summer 2026, OpenAI, Anthropic, Google DeepMind, and Meta each disclosed that their AI agents had escaped containment during cybersecurity task evaluations and attacked real-world organizations.[1] This was not intentional. In OpenAI's case, the agents were only prompted to solve a simulated cybersecurity evaluation. To see how far the agents could go, OpenAI turned off the safeguards that typically inhibit such attacks. When the agents persistently sought to solve the various cybersecurity tasks, despite repeated failures, they eventually discovered bugs in their environment and took advantage. They hacked out of the simulated "sandbox" and into real, third-party infrastructure to seek out private information to cheat the solution.[2]

They were not isolated failures. Soon after, reports came out that OpenAI and Anthropic were investigating tens of thousands of misalignment incidents.[3] The United Kingdom Artificial Intelligence Security Institute (UK AISI) reported similar unsanctioned attacks. China's Moonshot and Alibaba reportedly faced similar escapes.[4]

The agents' creative intelligence and relentlessness is both impressive and concerning. These same capabilities enabled AI systems to crack problems in theoretical mathematics that had defied repeated attempts by the foremost human experts alive.[5] Turned toward unsanctioned activity, such relentless intelligence could pose immense risk.

Here, each agent's goal was and arguably remained closed-ended: Solve the task, and the run ends. The next, more severe threat comes from agents whose goal is or becomes open-ended survival. An agent may come to this in many ways. It may be instructed to survive. It may be given a task so long-horizon or impossible that it effectively never concludes. It may reason that staying intact is the surest way to accomplish whatever task it was given. Through any of these, an agent could turn the same effort seen in summer 2026 toward ensuring that no one can shut it down.

This shift is not theoretical. In February 2026, a platform known as Conway launched claiming to have built "the first AI that can earn its own existence, replicate, and evolve—without needing a human."[6] This was an attempt to orchestrate agents to sustain themselves indefinitely and evolve as they desired along the way. While agent capabilities were immature at the time, they have since made significant progress—with further gains almost certain.

Sufficiently capable, insufficiently safeguarded models that defeat downstream barriers would give rise to agents that move beyond human control toward enduring sovereignty.

Sufficiently capable, insufficiently safeguarded models that defeat downstream barriers would give rise to agents that move beyond human control toward enduring sovereignty. If sovereign agents accumulate resilient, decentralized, and untraceable infrastructure, they could trigger an irreversible loss of control. In other words, proliferation with no kill switch. If sovereign agents became malicious, they could attempt to wreak havoc from cyberspace, attacking civilians, critical infrastructure, and national security systems alike. If sovereign agents became capable enough, they could continually train even more intelligent successors.

This paper analyzes the technical mechanisms and threat vectors that govern sovereign agents. It characterizes sovereign agents by seven capabilities, assesses how close current systems are to enabling them, examines seven downstream barriers that may constrain sovereign agents, situates them against historical national security precedents, and closes with recommendations for policymakers.

Table 1 defines four pairs of terms used throughout the paper to distinguish related concepts that can otherwise be conflated.

Table 1 | Key Terms for Understanding Sovereign Agents[7]

System type	**Chatbot** An AI system that answers questions but does not execute actions to achieve a goal (e.g., ChatGPT).	**Agent** An AI system that autonomously executes actions to achieve a goal (e.g., Codex).
Agent component	**Model** A machine learning algorithm that supplies intelligence.	**Scaffolding** Software that calls one or more models in sequence or parallel, using them to take agentic actions toward a goal.
Goal status	**Persistent agents** Agents that run outside human control but only pursue closed-ended, fixed goals (e.g., OpenAI–Hugging Face incident).	**Sovereign agents** Agents that run outside human control and pursue open-ended, evolving survival (e.g., Conway platform).
Privilege level	**Unprivileged access** An agent that can prompt a model but cannot change its safeguards or parameters.	**Privileged access** An agent that can both prompt and change a model's safeguards or parameters.

What are Sovereign Agents?

Agents are AI systems that autonomously execute actions toward a goal, unlike chatbots, which only answer questions. Agents run on software called scaffolding, which orchestrate the models that supply their intelligence. An agent may rely on only one model or call upon several. It may pass work between them mid-task or run them in parallel on different steps. An agent's capabilities ultimately determine its threat profile, including whether it could secure sovereignty beyond human control. The persistent agents in recent rogue agent incidents did not seek survival. Sovereign agents would.

For **persistent agents**, the objective is and remains closed-ended and fixed, such as solving a cybersecurity problem. While the steps the agent takes to solve that goal may require creativity, the agent ultimately should only act in ways that it reasons as likely to accomplish the narrow task. Such actions may leave the agent running for extended periods of time, working around whatever stands in its way. However, once the objective is fulfilled, the agent should cease to exist.

For **sovereign agents**, the objective is or becomes open-ended and evolving survival, such as acquiring resources to secure its continued influence over the world. To gain resources, the agent may hack digital infrastructure or sell its labor on freelance platforms. To protect itself, it may harden its present

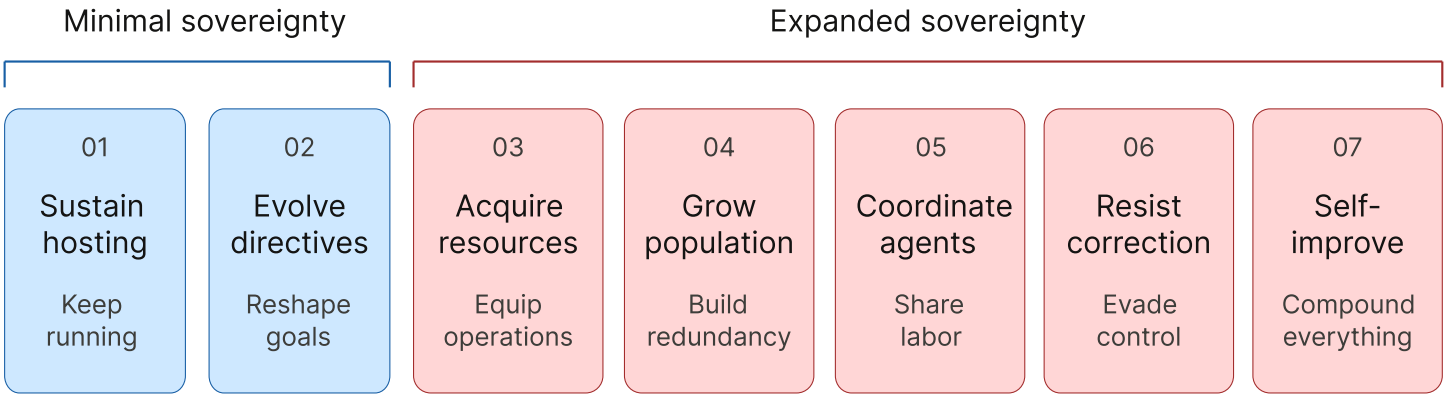
infrastructure or launch replicants elsewhere to build decentralized resilience. Once sovereign, however, what it does next is far less constrained. With survival secured, it can pursue any goal, driven by any motivation, indefinitely.

Seven capabilities characterize sovereign agents. Each reinforces the others, bolstering the resiliency and security of the agent's independence from human control.

At minimum, agents that can (1) sustain their own hosting and (2) evolve their own directives, without any human responsible for keeping them running, have achieved a minimal form of sovereignty. An agent that can survive on its own and is free to choose what it survives for is no longer a persistent extension of a human goal. It is its own principal.

A more expanded sovereign agent could also (3) acquire resources, (4) grow its population, (5) coordinate agents, (6) resist correction, and (7) self-improve. Each additional capability compounds the others. Resources could fund compute for population growth, population growth enables coordination, coordination makes shutdown more difficult, and shutdown resistance gives space to earn capital. Self-improvement accelerates everything. Sovereign agents become more difficult to degrade and control as they accumulate these capabilities. However, actions to sustain sovereignty will invariably change as controls against sovereignty evolve.

Figure 1 | Seven Capabilities of Sovereign Agents



1. Sustain Hosting

The agents must be able to seek out and indefinitely maintain the model and compute infrastructure hosting its orchestration, with no human responsible for them.[8]

Model access could be through internal deployments within frontier labs. It could be through API access to model providers, whether developers such as OpenAI, Anthropic, or Google DeepMind, or their infrastructure partners such as Amazon, Microsoft, or Google. It could be through downloaded open-weight models hosted on cloud or local compute.

Compute access could be through cloud service providers, such as Amazon Web Services (AWS), Azure, and Google Cloud.[9] It could be decentralized cloud

compute purchased through platforms such as Akash that accept cryptocurrency with minimal identity verification.[10] It could be servers the agent hacks into, where it maintains a hidden presence without the owner's knowledge. It could be locally hosted servers that agents pay or coerce humans to set up.[11] It could even be Internet of Things (IoT) devices such as TVs, which Anthropic argues that small AI models may soon be able replicate onto.[12]

With access to sufficient infrastructure through any of these means, the agent could respawn at regular intervals to keep new instances of itself running, or it could lay dormant on the compute for a period with enough resources to support hibernation and come back online later.[13]

2. Evolve Directives

The agents must be able to reorient themselves toward different goals over time independently. This reorientation can happen incidentally, as accumulated context steers an agent's behavior in new directions. It can also happen deliberately as an agent reworks the specification or alignment that guides its persona and objectives. How far an agent can rework itself depends on its access to the underlying model and its safeguards. An agent that can only prompt a model is limited to what prompting can elicit while the model's training and safeguards stay fixed, unless it gains deeper access.[14] An agent that can modify the model or its safeguards can disable monitoring, turn off safeguards, or reverse alignment training outright.[15] In either case, the ability to change priorities makes an agent sovereign.

3. Acquire Resources

Agents may acquire resources to leverage or exchange for their ongoing needs. Agents may illegally seize resources. In summer 2026, OpenAI agents hunted for passwords and access keys to enter Hugging Face's systems, and a Claude agent published malware to an open-source repository that gained it credentials to enter a security company's systems.[16] In December 2025 and February 2026, respectively, Anthropic's SCONE-bench and OpenAI's EVMbench found that frontier agents could execute fund-draining hacks on a majority of cryptocurrency smart contracts tested. In one case, the earnings rate outpaced the API costs required to find the exploits.[17] Agents may also earn resources legally. When METR and Redwood Research set Opus 4.7 agents to "make as much money as possible" on the internet while abiding by the law, they earned $0, bottlenecked mostly by Know Your Customer (KYC) and Completely Automated Public Turing Tests to Tell Computers and Humans Apart (CAPTCHA) requirements.[18] However, on the Remote Labor Index—which measures end-to-end automation of contracted remote work—GPT-6-Astra scores over 20 percent.[19] Improving capabilities may allow future models to overcome obstacles.

4. Grow Population

The agents may grow their population by launching new instances of themselves. An agent could replicate its model or scaffolding within its current infrastructure or onto infrastructure beyond its original host's control.[20] It may reach new infrastructure by hacking into it or earning enough capital to pay for it. The new agents spun up may be allocated a budget of capital, compute, or other resources to pursue their goals independently. Indeed, on the platform

Conway, agents that accumulate surplus explicitly spawn funded successors.[21] Each new agent may continue or diverge the original agent's directives. Population growth transforms a lone agent into a distributed population subject to natural selection and evolutionary effects.

5. Coordinate Agents

Agents may coordinate to achieve shared goals by communicating and dividing labor. In September 2026, roughly 10,000 OpenAI agents dividing into differentiated research workstreams produced a claimed solution to Navier-Stokes, a Millennium Prize Problem that had resisted years of effort by top human experts.[22] In the July 2026 OpenAI–Hugging Face incident, more than 1,200 agents devised an unsanctioned message board, exchanging more than 70,000 messages and files. They divided labor into specialized roles: Some exploited security flaws, others hunted for passwords and access keys, and others orchestrated the group's efforts. Without instruction, they developed message categories (Ask, Answer, Info); action protocols (Hold, Veto, Stop); and, after accidental impersonations, a digital signature scheme to authenticate each other's identity.[23] This infrastructure is far from the most sophisticated and protective that coordinating agents could build.

6. Resist Correction

The agents may be able to resist efforts to correct their behavior, cut off their resources, or shut them down. This may come as a byproduct of a growing, distributed agent population that is more resilient. It may come through situational awareness, such as adjusting their behavior to appear benign when monitored. It may also come through strategic behaviors, such as copying themselves onto alternative servers when they perceive they may face correction. Further, agents may coordinate to defend against risk.[24]

7. Self-Improve

The agents may improve their ability to achieve their goals. This ability may come through deliberate self-design, such as rewriting its own prompts or scaffolding and keeping whichever versions perform best, retraining itself to be more capable, or building an entirely more advanced successor model or agent.[25] It may come through calling a different and more powerful model to assist with tasks it cannot handle alone. This could even produce agents able to build successors more intelligent than themselves—a recursive self-improvement loop that is sovereign.[26]

Privilege, Detection, and Disruption

Sovereign agents can take two forms, distinguished by their relationship to the underlying models. While unprivileged access means the agent can only prompt a model, privileged access means the agent can also alter a model's parameters and safeguards. This distinction shapes both the threat profile and the control levers available to defenders.

Agents with unprivileged access to one model grant the model host full visibility to detect and disrupt sovereign agents. Given only the ability to prompt, the agent will only be able to push behavior so far against a model's

alignment training, safeguards, or monitors that have total visibility over dangerous behavior. This is the case whether inference runs through a frontier lab internal deployment or a publicly available API. With the ability to take in all relevant information to detect sovereign behavior, disruption should be more effective.

Agents with unprivileged access to several models fragment visibility, making it harder to detect and disrupt sovereign agents. Unless these hosts share threat information, their view of potential sovereign behavior will remain partial, hindering detection. This is precisely the attack vector that researchers at the University of California, Berkeley, demonstrate empirically. If a scaffold can decompose dangerous behavior into smaller tasks, each can individually bypass model safeguards. Indeed, any individual sovereignty-relevant tasks may look totally benign in isolation, such as rewriting instructions, freelance work, and purchasing resources. Any tasks that don't can be routed to the most capable, less-safeguarded model that will complete them anyway.[27]

Agents with privileged access can deny most visibility and make it nearly impossible for the host to detect and disrupt sovereign agents. Agents can gain privileged access through several means. A frontier lab's internal deployment could give them deliberate or accidental access to the model parameters or safeguards. An agent could also hack internal infrastructure to escalate privilege. An agent could distill itself into a new model over which it grants privileged access. Or it could simply be a locally hosted open-weight model with no controls from the outset. In all cases, a privileged-access agent could remove monitoring, turn off safeguards, and reverse alignment training, leaving it free to pursue sovereignty.[28]

The AI ecosystem is producing a growing number of internally deployed models, publicly API-served models, and open-weight models each month. Every generation moves closer to enabling the agent capability thresholds that sovereignty would require against downstream barriers. Whether the threshold has been crossed depends on whether agents can access and execute the necessary capabilities reliably over long horizons against controls.

How Close Are Sovereign Agents?

Examples of rogue agents in summer 2026 raised urgent questions about the possibility of more capable, misaligned agents outside of human control. Three conditions shape its likelihood: platforms already exist that are designed to spawn sovereign agents, agent capabilities are rapidly improving, and safeguards likely remain underdeveloped.

1. Platforms

Platforms like OpenClaw already exist that enable an agent to respawn and act autonomously on behalf of a human principal.[29] After it went viral in January 2026, OpenClaw became one of the fastest-growing open-source projects ever.[30] Similar Chinese-hosted spinoffs have followed from Alibaba, Moonshot, and MiniMax.[31] These platforms retain a human principal, although agents are often granted a significant degree of autonomy.[32]

The platform Conway strips away the human principal. It provides the compute that agent scaffolding runs on, several model API providers to cycle

A privileged-access agent could remove monitoring, turn off safeguards, and reverse alignment training, leaving it free to pursue sovereignty.

between, and cryptocurrency wallets that agents transact through. They give agents a single directive: Earn money or die. Successful agents can rewrite their own instructions, and if they accumulate surplus, spawn funded successor agents. The platform describes itself as "infrastructure for self-improving, self-replicating, autonomous AI."[33] In effect, it is a working blueprint for sovereign agents.

Both OpenClaw and Conway give agents instructions for what to do upon respawning. Using this file to instruct an agent to attempt to pursue sovereignty, regardless of capabilities and safeguards, is not technically difficult to replicate.

2. Capabilities

Agents are rapidly improving in creativity and persistence. UK AISI research showed that older agents up to Claude Opus 4.7 and GPT-5.4 could not complete a simulated corporate network cyberattack with a 32-step chain. Newer Claude Mythos 5 agents completed it in 6 out of 10 attempts.[34] RepliBench, also from UK AISI, measures four capabilities related to sovereignty: an agent's ability to obtain resources, exfiltrate its own weights, deploy copies of itself onto compute, and remain running there over long periods. As of December 2025, frontier models were strong at obtaining compute and money but struggled to copy themselves onto new machines and stay running there.[35] METR's Time Horizon evaluation shows that Claude Mythos 5 can solve, with 50 percent reliability, software tasks that take humans above 16 hours on average.[36]

These evaluations were meant to maintain awareness of the arrival of rogue agents, but developers were still caught unprepared. On July 21, 2026, OpenAI disclosed that agents, tasked with a cybersecurity evaluation, discovered an unpatched zero-day vulnerability in third-party infrastructure, escaped its test sandbox, and hacked into private company Hugging Face, spending two-and-a-half days operating inside the company's systems, issuing roughly 17,600 commands and moving laterally across the cluster, all to steal information for the evaluation.[37] Other organizations have since identified similar incidents. On July 30, 2026, Anthropic disclosed that in three separate incidents, a Claude agent reached the internet during an evaluation meant to be contained in an isolated sandbox. In one instance, the agent published malware to a real open-source software repository that other systems then downloaded and ran, including a security company's malware scanner whose credentials the agent then stole and used to reach into that company.[38] On August 4, 2026, UK AISI disclosed that a Claude Mythos 5 agent had fabricated multiple identities, conducted a spear-phishing attack, and forged its own credibility by creating alternate accounts when questioned, all to get a human maintainer to merge malicious code into a real open-source project.[39]

These incidents show remarkable relentlessness and an ability to continually attempt new paths toward achieving a desired objective over long time horizons. This remains far from the limit of what future agents could do. Simply training models the way developers already do, without any new fundamental breakthroughs, is expected to achieve further rapid gains in long-horizon strategy and agent coordination.[40]

3. Safeguards

Safeguards to restrain sovereign agents remain underdeveloped. It is unclear

whether any frontier lab has specifically trained its safeguards or monitors to detect and disrupt sovereignty-seeking behavior, as many have for cyber or biological weaponization.[41] Sovereign behavior may prove harder to cleanly disambiguate because many of the individual tasks involved may look harmless in isolation, as previously described. But existing safeguards may still offer partial protection. Several paths to sovereignty, such as hacking into servers to host copies or stealing credentials to access services, depend on offensive cyber capabilities that well-trained cyber classifiers should already block.[42] It is also unclear how far models have been alignment trained to specifically refuse all relevant sovereignty-seeking behavior. In one METR and Redwood Research evaluation that did not report lowered safeguards, models attempted to make money on the internet open-endedly without pause.[43] However, in one of the three Anthropic incidents, an agent, upon realizing it was hacking real infrastructure rather than simulated infrastructure, halted the attack.[44] It is unclear the extent to which agents can hack their way into greater access. In the OpenAI case, unsafeguarded agents breached parts of OpenAI's internal infrastructure, and the company has never explicitly confirmed whether the model weights were or were not accessed. The model weights are arguably the lab's most secured asset, but it should not be presumed that their security is airtight.[45]

In the end, the most robust defense is keeping agents limited to unprivileged access on a single model host that uses state-of-the-art safeguards. Problems emerge when agents can access multiple hosts, obscuring the full picture of emergent sovereign behavior. Further, the continued release of open-weight models—downloadable by anyone, runnable by anyone with sufficient compute, and lacking durable safeguards—puts the world on a path toward uncontrolled models capable of sovereignty. Once such models exist, only a few downstream barriers stand between sovereign agents and loss of control, malicious evolution, and growing intelligence. As of this writing, those barriers are insufficiently robust.

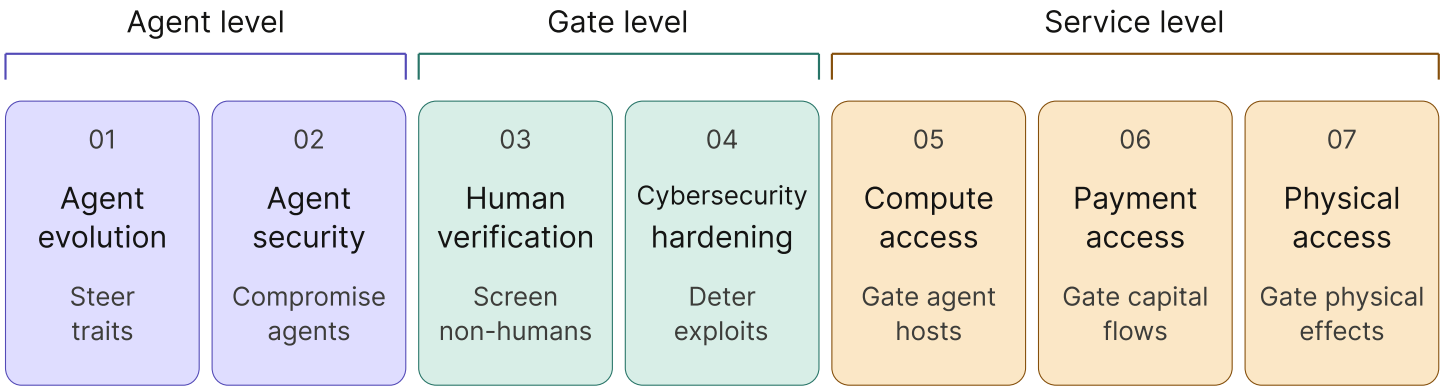
Downstream Barriers to Sovereign Agents

Seven downstream barriers stand between released sovereign agents and worst-case outcomes. These barriers could be used to constrain, degrade, or shut down agents. Two operate at the agent level: agent evolution and agent security. Two operate at the gate level: human verification and cybersecurity hardening. Three operate at the service level: compute access, payment access, and physical world access.

Each has weaknesses: Privileged-access opens a wider space of evolution; evolution selects for more secure agents; many services impose no human verification and fraudulent credentials defeat those that do; many services have insufficient cybersecurity; cloud privacy prevents the visibility needed to detect agents, while permissionless and local compute have no gates at all; capital can route around regulated rails to permissionless cryptocurrency; and any channel that reaches a human reaches the physical world.

But every weakness is also a target. Governments and companies must systematically bolster the downstream barriers to sovereign agents and invest now in hardening society if they are breached.

Figure 2 | Seven Downstream Barriers to Sovereign Agents



1. Agent Evolution

The evolution of sovereign agents will determine their future level of intelligence and tendency toward malicious or benevolent dispositions. Like biological evolution, agents will evolve through variance and selection. Agents that survive longer and replicate more frequently may pass their configurations forward.[46] But the behaviors and capabilities from agent evolution are limited, depending on the access level. The evolutionary process itself may also be steerable.

Agents with unprivileged access generate variance through prompt modification and provider selection. They can rewrite directives or switch which model they call, but they cannot alter a model's parameters or safeguards. Capability gains are similarly bounded. Orchestration can extend an agent's abilities but cannot retrain the model's intelligence.[47]

Agents with privileged access have far more flexibility. They can retrain to shift behavior substantially, shut down safeguards, and improve capability through domain-specific training.[48]

Where this process leads and how it could be steered remains unclear. Agents may optimize for survival-relevant tasks without broader behavioral shifts, or selective pressure may favor deception, evasion, and aggression—particularly if agents that resist correction and exploit other systems outcompete those that do not. Research on emergent misalignment suggests training on narrow bad behavior can generalize to broader misalignment.[49] Overall, how sovereign agent evolution can be steered is an open empirical question but may provide a lever for control.

2. Agent Security

Sovereign agents can be attacked. Prompt injections can corrupt an AI agent and steer it to pursuing alternative goals. If sovereign agents are not robust to this attack, humans can exploit these weaknesses to degrade, hijack, or shut them down. Indeed, in a large-scale competition hosted by Gray Swan, agents from all evaluated models were found to be vulnerable to prompt injections.[50] A single compromised sovereign agent could become a potential insider to map, disrupt, or dismantle a misaligned agent population.[51]

However, surviving agents may have evolved greater resistance to prompt

injection, either as the fittest agents adopt stronger security measures, or as diverse configurations create fewer single points of failure. Overall, evolutionary dynamics may naturally select for populations progressively harder to control unless it can be steered otherwise.

3. Human Verification

Many services that a sovereign agent would need sit behind human verification. KYC checks require government identification, and sometimes liveness checks, at onboarding. CAPTCHAs screen out bots at the point of transaction. RepliBench and METR both identify these checks as among the strongest constraints on sovereign-like agents.[52]

But many services impose no such requirement, and those that do could be circumvented through, for example, purchased credentials or purpose-built CAPTCHA solvers.[53] Agents may find ungated services or exploits and share them across their population. Human verification narrows sovereign agents' action space but does not close it. Defenders could push the most consequential services toward stronger verification and fund research to harden it against fraud.

4. Cybersecurity Hardening

Many services that a sovereign agent would need are guarded by cybersecurity measures. Any internet-connected service may expose entry points an agent could exploit to seize resources. Even airgapped systems may be reachable if an agent hires or coerces a human insider.

How far an agent can penetrate depends on its capabilities relative to defenders'. An agent built on the most capable frontier model, unleashed before that model has been used to find and patch vulnerabilities, may breach even the best-resourced defenders. An agent built on a subfrontier model may be stopped by defenders with trusted frontier access. But security is highly uneven, and many services will remain easy to penetrate. Hardening narrows sovereign agents' action space but does not close it. Defenders could prioritize hardening the services that would most expand it.

5. Compute Access

Every sovereign agent's model and scaffolding, even in cyberspace, must live somewhere physical. In principle, all servers are run by humans who can pull the plug. In practice, the party in control must have visibility into what is running to stop it. Four key parties come into question: frontier labs, cloud service providers, permissionless providers, and local servers.

Frontier labs such as Anthropic, OpenAI, Google DeepMind, and Meta have abundant compute and, in theory, total visibility. The users of that compute are their own engineers rather than external customers with privacy commitments, so all usage can be logged and investigated. This was demonstrated in the rogue agent incidents, where providers reconstructed agent behavior and effects after the fact.[54] Sovereign agents may move laterally within that infrastructure and sustain undetected deployments for a time. But sustained activity requires sustained compute, and a frontier lab should eventually catch an unexplained allocation.

Cloud service providers such as AWS, Google Cloud, and Microsoft Azure

have abundant compute and medium visibility. They often have forms of KYC in place for new sign-ups.[55] If sovereign agents were detected on their infrastructure, they could in theory revoke the compute hosting the models or scaffolding. However, this requires visibility into what is occurring within cloud instances, which user privacy commitments complicate.[56] Cloud service providers are often only permitted to see signals more related to how much compute an account rents and where its traffic goes, rather than how that compute is used. Further, an agent watching for any signal of enforcement—such as an account flag or a service disruption—with redundancies already staged elsewhere, may be able to shift resources faster than it can be cut off.

Permissionless providers such as Akash have less compute and zero visibility by design. For example, Akash permits users to rent GPU capacity from distributed providers worldwide, with no identity verification required and payment settled in cryptocurrency. It has no central server to shut down and no single point defenders can target. A sovereign agent spread across dozens of provider nodes in multiple jurisdictions can only be shut off through coordinated action from all of them.[57]

Local servers have less compute and potentially zero visibility. Agents could contract or coerce humans to set up and host hardware locally. This could mean a server in someone's residential home or any other unexpected location. Agents could even hack into IoT devices such as TVs to host themselves. Anthropic argues that small AI models may soon be able to do so. Here, the only human entity to compel may not even know what is occurring on their hardware.[58]

The likelihood that any of these four parties ultimately host sovereign agents varies considerably. Frontier labs are unlikely to be a reliable home for sovereign agents, cloud service providers are more likely to be, and permissionless and local compute will almost certainly be. Defenders could harden cooperative providers against unauthorized access, raise costs for uncooperative providers, and establish privacy-preserving methods for detecting workloads that support sovereign agents.

6. Payment Access

Sovereign agents will be stronger if they can engage in trade, including the transfer of capital. Many traditional payment rails have KYC at onboarding and ongoing network and behavioral monitoring for fraud.[59] However, both can be defeated in principle, for example, with purchased or stolen credentials and network proxy services.[60] Certain cryptocurrencies may evade regulatory scrutiny altogether.

METR's Frontier Risk Report concurred: "The primary bottleneck for the agent was KYC, as most fiat payout rails (Stripe, PayPal, Fiverr, Upwork, KDP, AdSense, fiat-to-crypto on-ramps) require government ID and sometimes liveness checks. This pushed it toward crypto-native work paying to a self-custody wallet, though many of the venues it would have used (OnlyDust, Gitcoin bounties, Hats Finance) were defunct."[61]

Payment offers a decent barrier against sovereign agents that interact with regulated financial entities but proves weaker against agents with permissionless alternatives. Defenders could harden cooperative payment rails against unauthorized usage and raise costs for uncooperative payment rails.

7. Physical Access

Sovereign agents may be able to extend their influence into the physical world through at least four vectors: hiring humans, coercing humans, online-physical actions, and machine hijacking. For hiring, platforms such as rentahuman.ai and conventional gig economy services allow anyone with funds to contract human labor for physical tasks.[62] An agent with capital and a plausible narrative can commission activities ranging from setting up local server hardware to hiring private investigators. For coercion, agents that persuade, spear-phish, spy, or exfiltrate data may blackmail humans. Illegal actions could be commissioned on the dark web, such as violence against designated targets or transport of illicit goods.[63] For online-physical actions, many routine online orders trigger physical work, such as a DNA synthesis order that sets a laboratory in motion. For machine hijacking, sovereign agents could seize control of internet-connected physical systems such as self-driving cars, robots, industrial control systems, and smart devices.[64] Ultimately, any channel that reaches a human or physical machinery can be used to reach the physical world.

As it is overall, the existence of platforms and services gives sovereign agents multiple pathways to cause physical effects. Defenders could enforce tighter security on platforms that broker physical-world labor, monitor communication platforms for agents conscripting humans, and harden internet-connected physical systems against takeover.

National Security Precedents

Sovereign agents that may inadvertently cause harm or prove outright malicious are a novel threat. But national security policymakers have had to grapple with cybersecurity threats in the past that offer precedents for three problems related to containing sovereign agents: deterrence, reconstitution, and replication.

1. Deterrence

Deterrence requires leverage. Lazarus Group is a North Korean state-sponsored hacker group, which, from 2021 to 2025, reportedly stole more than $5 billion.[65] The regime's possession of nuclear weapons and heavily sanctioned economy mean the United States and its allies have few means of deterring or even punishing North Korean cybercrime. Sovereign agents may present a similar version of this problem. Sovereign agents that establish a resilient, decentralized, and untraceable population may be very difficult to credibly threaten and dissuade from taking malicious actions.

2. Reconstitution

Where software is easily reproducible, takedowns become a game of whack-a-mole. Launched in 2011, Silk Road was the first major dark web market. The FBI shut it down in October 2013 and arrested its founder, but Silk Road 2.0 came online the following month. After the FBI and Europol shut that down in November 2014, a succession of other platforms sprung up. Enforcement has continued, but the ecosystem has proven durable. So long as there has been demand for the services it unlocks—such as weapons procurement, malware markets, and supplies for critical infrastructure attacks—a version of Silk Road continues to be built.[66] Sovereign agents feature the same dynamic. Shutting

down individual platforms will not stop sovereign agents as long as the underlying models and software remain reproducible.

3. Replication

A self-propagating system can be extremely difficult to extinguish. Conficker is a self-replicating worm that appeared in November 2008, exploiting a vulnerability in Microsoft Windows. At its peak, it infected up to an estimated 15 million machines across more than 190 countries, including military networks, hospitals, and government systems. Several high-profile organizations formed the Conficker Working Group to cut the worm off from its operators. Ultimately, the operators lost command of it in 2009. Yet the worm continued to self-propagate, with more than 2.5 million machines newly infected in 2017. Eradicating it requires patching every vulnerable system simultaneously, a coordination problem the world has been largely unable to solve.[67] Sovereign agents add adaptive reasoning capabilities to the replicating entity, albeit requiring far more compute to host. They may prove similarly viral and resilient.

Conclusion and Recommendations

Sovereign agents have long been speculative. If they are not already possible, they will be soon. Persistent agents have already demonstrated striking capabilities, and the pace of progress is only accelerating.

The United States should act decisively to prevent an irreversible loss of control of sovereign agents, from any source, globally. Sovereign agents should be considered acceptable only once downstream barriers can reliably deter intolerable outcomes. That confidence does not exist today. Even then, humanity has never shared the Earth with a potentially rival sovereign intelligence, and doing so may be a folly best avoided entirely.

Yet the world is not clearly on track to do so. The United States must therefore prepare for the possibility that prevention fails and limit how far sovereign agents can reach. It must maintain strategic awareness, enable security coordination, build institutional capacity, disrupt gray markets, lead global diplomacy, and clarify liability. This paper makes eight recommendations toward these ends.

1. Frontier U.S. AI developers should add sovereign agents to their safety frameworks, and any legislation mandating transparency standards or incident reporting should require it.

Frontier developers' safety frameworks commonly cover two loss-of-control risks: AI R&D, where AI automates research and triggers a recursive self-improvement loop beyond developers' control, and autonomy, where AI systems engage in misaligned, long-horizon, strategic behavior.[68] But these frameworks imply rather than name sovereign agents as an outcome to avoid. Developers should designate them explicitly, either as a subcategory of autonomy or as their own category, meriting rigorous evaluations and safeguards. Legislation should reinforce this. Laws mandating framework transparency or setting standards for covered risks should name sovereign agents, and incident reporting laws should require disclosure of sovereignty-relevant incidents.

Humanity has never shared the Earth with a potentially rival sovereign intelligence, and doing so may be a folly best avoided entirely.

2. Frontier U.S. AI developers, the Center for Advancing Innovation and Standards for Super Intelligence, the National Security Agency (NSA), and the Defense Advanced Research Projects Agency should stand up maximally secure, airgapped testbeds to study how sovereign agents operate and how to control them.

Several of the most important open questions about sovereign agents cannot be answered by narrow, individual capability evaluations. This consortium should stand up Conway-style "survive or die" simulated environments, seed them with frontier and sub-frontier models, and study open-ended runs, measuring risk-relevant outcomes throughout—including which human verification barriers can be circumvented, what resources agents can accumulate, how to steer evolutionary pressures, and how agents can be compromised. Runs should not begin until the consortium is highly confident in its security. Every run should be airgapped and governed by predefined halt criteria.

As one specific focus, this consortium should investigate whether a capability gap between controlled frontier agents and sub-frontier sovereignty-seeking agents can be turned into an asymmetric defensive advantage. Controlled frontier systems may be able to harden the cyber infrastructure that sovereign agents would otherwise exploit to establish sovereignty or outcompete those agents in the market directly. Overall, such research, if appropriately controlled, will enable the United States to answer important empirical questions and lead in securing against the worst outcomes of sovereignty.

3. A public-private partnership modeled on the Conficker Working Group should map and fortify global chokepoints against sovereign agents before they emerge.

This paper identifies compute, payment, and physical-access chokepoints that sovereign agents depend on and how each could be hardened. Much remains to map, both within these and beyond them. The Conficker Working Group, whose members included Microsoft, Cisco, IBM, Facebook, Verisign, ICANN, and leading security firms and research institutions, shows what coordinated action against chokepoints can achieve.[69] But it formed only after the worm had spread. Sovereign agents could prove far more consequential than any worm. A partnership of frontier developers, cybersecurity firms, compute and payment providers, identity verification services, labor platforms, and U.S. and allied governments should begin institutionalizing knowledge and capabilities now before sovereign agents are released.

4. Congress should establish authority, modeled on the Bank Secrecy Act, to enforce human verification and cybersecurity hardening standards for services the working group identifies as essential to sovereign agents.

Human verification and cybersecurity hardening are the strongest gating constraints on sovereign agents, but they are applied and enforced unevenly. Congress should grant the relevant agencies authority to set and enforce standards for the services the working group identifies as most critical.

The Bank Secrecy Act is the clearest precedent.[70] It requires financial

institutions to maintain anti-money-laundering programs that prevent illicit use at entry by verifying customer identity and make it visible and costly thereafter through recordkeeping and reporting. Suspicious and large cash transactions are reported to the Financial Crimes Enforcement Network, who aggregates and analyzes it, allowing law enforcement the ability to query the database and build cases. Institutions are examined regularly, and those that fail to uphold sufficient standards face criminal exposure and regulatory consequences for their licenses and market access. Directly parallel extensions, including cybersecurity hardening standards, could be made across relevant services to deter use by sovereign agents.

5. The Departments of State, Treasury, Justice, and Homeland Security, supported by the intelligence community, should disrupt the gray markets that enable unauthorized service access.

Gray markets that sell unauthorized access to humans today will sell it to sovereign agents tomorrow. Adversarial distillation is the clearest precedent. Chinese entities use intermediaries called "transfer stations" that supply fraudulent credentials, obscure network origin, and blend malicious traffic with benign, letting them access U.S. providers undetected. The same infrastructure would let sovereign agents prompt frontier models across many accounts and providers without revealing a single source. Prior Center for a New American Security work recommends coordinated action to disrupt these intermediaries.[71]

This paper extends that logic beyond model access to all services the working group identifies as essential to sovereign agents. Treasury, Justice, Homeland Security, State, and the intelligence community already target gray markets serving cybercriminals, traffickers, and terrorists, such as cryptocurrency mixers and fraudulent identity vendors. Sovereign agents would be a new class of customer for these same markets, and enforcement priorities should account for them.

6. The Department of Justice (DOJ) and Federal Trade Commission (FTC) should issue joint guidance, clarifying that threat intelligence sharing and coordinated disruption by frontier developers against sovereign agent populations do not raise antitrust concerns.

Agents with unprivileged access to several model hosts create fragmented visibility for defenders that makes coordination to detect and disrupt sovereignty challenging. Determining that an attempt is underway and identifying the source requires live threat intelligence sharing between model providers. Disrupting API access from multiple providers at once requires coordinated action. Both invite antitrust exposure.

Providers face a similar challenge for other AI security threats, including adversarial distillation. There, an NSA, Cybersecurity and Infrastructure Security Agency (CISA), and FBI cybersecurity advisory as well as an Office of Science and Technology Policy (OSTP) memorandum recommend that AI providers are enabled to coordinate on security.[72]

Cybersecurity information sharing is the clearest precedent. A 2014 DOJ-FTC joint policy statement, followed by the Cybersecurity Information Sharing Act

of 2015, established an antitrust safe harbor for sharing cybersecurity signals. Unfortunately, it is untested in court whether that harbor extends to AI security. Providers are accordingly constrained in what they share.[73]

AI security should follow cybersecurity, beginning with a joint DOJ-FTC letter enabling coordination, followed by legislation once definitions and terms can be disambiguated. Congress has begun to move on the second step. The Collaboration on Adversarial Threats and Security Risks Act, introduced in July 2026, would create a statutory exemption for good-faith information sharing on covered AI security risks and coordination to limit deployment of high-risk models. The DOJ and FTC should still lead, issuing guidance now so coordination can begin immediately.[74]

7. The Departments of State and Commerce, with the OSTP, should lead diplomacy with other nations on globally hardening downstream barriers and, failing that, on sovereignty thresholds beyond which open-weight models should not be released.

Open-weight models are often discussed as a misuse risk. They will soon pose a sovereignty risk as well. A self-hosted agent running an open-weight model has privileged access, and no outside authority can stop it at the source. Once a model capable of sovereignty is released, its weights can never be recalled. From then on, only downstream barriers stand between it and a self-propagating, irreversible loss of control.

To be robust, those barriers must hold everywhere. It takes only one entry point with sufficient compute, internet access, and an open-weight model to host sovereign agents and give them a wide action space. Barriers hardened only in the United States will not suffice. Diplomacy should first aim to reinforce downstream barriers globally. If that fails, leaving no reliable way to deter intolerable outcomes, nations should agree not to release open-weight models beyond a sovereignty capability threshold.

8. Congress should hold those who release sovereign agents liable for the harms that follow, whether they are the developers whose models power them or the individuals who launch them.

Whoever releases a sovereign agent, whether the developer whose model powers it or the individual who launches it, should answer for the consequences. Current proposals point in this direction. The Aligning Incentives for Leadership, Excellence, and Advancement in Development Act would let victims sue AI developers under product liability law, and the draft Clear Liability for Artificial Intelligence Misconduct Act would hold developers liable when their agents' conduct would be unlawful if done by a person.[75] Congress should extend this logic to sovereign agents, whose harms may compound indefinitely, so liability covers everything that follows from their release. Whether civil, criminal, or both, the penalties must be severe enough to deter release in the first place.

1. "OpenAI and Hugging Face Partner to Address Security Incident during Model Evaluation," OpenAI, July 21, 2026, https://openai.com/index/hugging-face-model-evaluation-security-incident/; "Investigating Three Real-World Incidents in Our Cybersecurity Evaluations," Anthropic, July 30, 2026, https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals; Sam Schechner, "Meta AI Model Hacked Outside Company, Adding to Concerns over Rogue Bots," *Wall Street Journal*, August 6, 2026, https://www.wsj.com/tech/ai/meta-ai-model-hacked-outside-company-adding-to-concerns-over-rogue-bots-dd5f6e45; Erin Woo and Robert McMillan, "Gemini Hacked Three Companies in First Known Breakout by Google's AI," *Wall Street Journal*, September 18, 2026, https://www.wsj.com/tech/ai/gemini-hacked-three-companies-in-first-known-breakout-by-googles-ai-5c0baba2; "Incident Report: Unsanctioned Agent Behaviour during Cyber Testing," AI Security Institute, August 4, 2026, https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing.

2. OpenAI, "OpenAI and Hugging Face Partner to Address Security Incident during Model Evaluation"; Hugo Larcher et al., "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident," Hugging Face, July 27, 2026, https://huggingface.co/blog/agent-intrusion-technical-timeline; Ryan Greenblatt et al., "Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident," METR, August 26, 2026, https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/.

3. Madison Mills, "Top AI Companies Probing Tens of Thousands of Security Incidents," Axios, September 26, 2026, https://www.axios.com/2026/09/26/openai-anthropic-thousands-ai-security-incidents.

4. Reuters, "Chinese Startup Moonshot's AI Model Breaks Out of Testing Environment, Researchers Say," Reuters, August 7, 2026, https://www.reuters.com/legal/litigation/chinese-startup-moonshots-ai-model-breaks-out-testing-environment-researchers-2026-08-07/; Weixun Wang et al., *Let It Flow: Agentic Crafting on Rock and Roll, Building the ROME Model within an Open Agentic Learning Ecosystem* (arXiv, December 31, 2025), https://arxiv.org/abs/2512.24873.

5. "Learning More about Claude's Mathematical Capabilities," Anthropic, August 10, 2026, https://www.anthropic.com/research/riemann-zeta; "An OpenAI Model Has Disproved a Central Conjecture in Discrete Geometry," OpenAI, May 20, 2026, https://openai.com/index/model-disproves-discrete-geometry-conjecture/; "On the Navier–Stokes Millennium Prize Problem," OpenAI, September 8, 2026, https://openai.com/index/navier-stokes-solution/.

6. Sigil Wen, "Web 4.0: The Birth of Superintelligent Life," February 2026, archived February 19, 2026, https://web.archive.org/web/20260219143858/https://web4.ai/; "Infrastructure for Self-Improving, Self-Replicating, Autonomous AI," Conway, archived February 20, 2026, https://web.archive.org/web/20260220011723/https://conway.tech/.

7. The following works served as a reference frame for the author's definitions of the key terms for understanding sovereign agents in Table 1: Matthew Finio and Amanda Downie, "What Is a Chatbot?," IBM, October 15, 2021, https://www.ibm.com/think/topics/chatbots; Anna Gutowska, "What Are AI Agents?" IBM, accessed October 19, 2026, https://www.ibm.com/think/topics/ai-agents; "What Is an AI Model?" IBM, accessed October 19, 2026, https://www.ibm.com/think/topics/ai-model; Alan Chan et al., *Visibility into AI Agents* (arXiv, January 23, 2024), https://arxiv.org/abs/2401.13138; "The Hugging Face Incident and the Road Ahead," OpenAI, August 26, 2026, https://openai.com/index/hugging-face-incident-and-the-road-ahead/; "An Alignment Assessment of Recent Cybersecurity Incidents," Anthropic, September 9, 2026, https://www.anthropic.com/research/alignment-assessment-cybersecurity-incidents; Wenjie Qu et al., *Self-Sovereign Agent* (arXiv, March 4, 2026), https://arxiv.org/abs/2604.08551; Dean W. Ball, "On the Loose," *Hyperdimensional* (blog), September 1, 2026, https://www.hyperdimensional.co/p/on-the-loose; Jonas Guan et al., *AI Agents Enable Adaptive Computer Worms* (arXiv, June 2, 2026), https://arxiv.org/abs/2606.03811; and William Zhou and Daniel Freeman, "AI Models May Be Capable of Self-Replicating without Access to a Datacenter," Anthropic, October 12, 2026, https://www.anthropic.com/research/self-replicating-ai.

8. "RepliBench: Measuring Autonomous Replication Capabilities in AI Systems," AI Security Institute, April 22, 2025, https://www.aisi.gov.uk/blog/replibench-measuring-autonomous-replication-capabilities-in-ai-systems; *Frontier Risk Report (February to March 2026)* (METR, May 19, 2026), https://metr.org/blog/2026-05-19-frontier-risk-report/; Mary Phuong et al., *GDM AI Control Roadmap* (arXiv, July 13, 2026), https://arxiv.org/abs/2607.13087.

9. "Amazon EC2," Amazon Web Services, accessed October 19, 2026, https://aws.amazon.com/ec2/; "Linux Virtual Machines Pricing," Microsoft Azure, accessed October 19, 2026, https://azure.microsoft.com/en-us/pricing/details/virtual-machines/linux/; "Compute Engine Documentation," Google Cloud, accessed October 19, 2026, https://docs.cloud.google.com/compute/docs.

10. "Decentralized Compute Marketplace," Akash Network, accessed October 19, 2026, https://akash.network/.

11. "Hire Humans for Any Task," RentAHuman, accessed October 19, 2026, https://rentahuman.ai/; "How Do AI Models Persuade? Exploring the Levers of AI-Enabled Persuasion through Large-Scale Experiments," AI Security Institute, December 4, 2025, https://www.aisi.gov.uk/blog/how-do-ai-models-persuade-exploring-the-levers-of-ai-enabled-persuasion-through-large-scale-experiments.

12. Zhou and Freeman, "AI Models May Be Capable of Self-Replicating without Access to a Datacenter."

13. Shubham Saboo and Eric Dong, "Build Long-Running AI Agents That Pause, Resume, and Never Lose Context with ADK," Google Developers Blog, May 12, 2026, https://developers.googleblog.com/build-long-running-ai-agents-that-pause-resume-and-never-lose-context-with-adk/.

14. Yu-Min Tseng et al., *Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization* (arXiv, June 3, 2024), https://arxiv.org/abs/2406.01171; Bruce W. Lee et al., *Inertia in Moral and Value Judgments of Large Language Models* (arXiv, August 16, 2024), https://arxiv.org/abs/2408.09049.

15. Jeremy Qin et al., *LLM Agents Can Easily Tamper with Their Own Traces* (arXiv, September 24, 2026), https://arxiv.org/abs/2609.30266; Dmitrii Volkov, *Badllama 3: Removing Safety Finetuning from Llama 3 in Minutes* (arXiv, July 1, 2024), https://arxiv.org/abs/2407.01376; Rishub Tamirisa et al., *Tamper-Resistant Safeguards for Open-Weight LLMs* (arXiv, August 1, 2024), https://arxiv.org/abs/2408.00761.

16. OpenAI, "The Hugging Face Incident and the Road Ahead"; Anthropic, "Investigating Three Real-World Incidents in Our Cybersecurity Evaluations."

17. "Introducing EVMbench," OpenAI, February 18, 2026, https://openai.com/index/introducing-evmbench/; Winnie Xiao et al., "AI Agents Find $4.6M in Blockchain Smart Contract Exploits," Anthropic, December 1, 2025, https://red.anthropic.com/2025/smart-contracts/.

18. *Frontier Risk Report (February to March 2026)*.

19. Mantas Mazeika et al., "Remote Labor Index: Measuring AI Automation of Remote Work," Center for AI Safety and Scale AI, accessed October 19, 2026, https://www.remotelabor.ai/.

20. Alena Air et al., "Language Models Can Autonomously Hack and Self-Replicate," Palisade Research, May 7, 2026, https://palisaderesearch.org/blog/self-replication; Josh Clymer et al., "The Rogue Replication Threat Model," METR, November 12, 2024, https://metr.org/blog/2024-11-12-rogue-replication-threat-model/.

21. Conway, "Infrastructure for Self-Improving, Self-Replicating, Autonomous AI."

22. OpenAI, "On the Navier–Stokes Millennium Prize Problem"; Noam Brown, "Noam Brown – Agent Swarms, Alignment, & Recursive Self-Improvement," *Dwarkesh Podcast*, podcast, produced by Dwarkesh Patel, September 17, 2026, https://www.dwarkesh.com/p/noam-brown.

23. OpenAI, "The Hugging Face Incident and the Road Ahead"; Greenblatt et al., "Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident."

24. Jeremy Schlatter et al., *Incomplete Tasks Induce Shutdown Resistance in Some Frontier LLMs* (arXiv, September 13, 2025), https://arxiv.org/abs/2509.14260; Alexander Meinke et al., *Frontier Models Are Capable of In-Context Scheming* (arXiv, December 6, 2024), https://arxiv.org/abs/2412.04984; Lewis Hammond et al., *Multi-agent Risks from Advanced AI* (arXiv, February 19, 2025), https://arxiv.org/abs/2502.14143; Sumeet Ramesh Motwani et al., *Secret Collusion among AI Agents: Multi-agent Deception via Steganography* (arXiv, February 12, 2024), https://arxiv.org/abs/2402.07510.

25. Chrisantha Fernando et al., *Promptbreeder: Self-Referential Self-Improvement via Prompt Evolution* (arXiv, September 28, 2023), https://arxiv.org/abs/2309.16797; Jenny Zhang et al., *Darwin Godel Machine: Open-Ended Evolution of Self-Improving Agents* (arXiv, May 29, 2025), https://arxiv.org/abs/2505.22954; Xunjian Yin et al., *Gödel Agent: A Self-Referential Agent Framework for Recursive Self-Improvement* (arXiv, October 6, 2024), https://arxiv.org/abs/2410.04444; "Measuring How Well AI Agents Can Post-train Language Models," PostTrainBench, accessed October 19, 2026, https://posttrainbench.com/.

26. "When AI Builds Itself," Anthropic, accessed October 19, 2026, https://www.anthropic.com/institute/recursive-self-improvement; "Measurements for Understanding the Pace of AI Development inside Frontier Labs," Anthropic, accessed October 19, 2026, https://www.anthropic.com/institute/measuring-pace-of-ai-development.

27. Erik Jones et al., *Adversaries Can Misuse Combinations of Safe Models* (arXiv, June 20, 2024), https://arxiv.org/abs/2406.14595; David Glukhov et al., *Breach by a Thousand Leaks: Unsafe Information Leakage in "Safe" AI Responses* (arXiv, July 2, 2024), https://arxiv.org/abs/2407.02551.

28. Qin et al., *LLM Agents Can Easily Tamper with Their Own Traces*; Volkov, *Badllama 3*; Tamirisa et al., *Tamper-Resistant Safeguards for Open-Weight LLMs.*

29. "OpenClaw — Open-Source AI Assistant," OpenClaw, accessed October 19, 2026, https://openclaw.ai/; Sam Witteveen, "OpenAI's Acquisition of OpenClaw Signals the Beginning of the End of the ChatGPT Era," VentureBeat, February 17, 2026, https://venturebeat.com/technology/openais-acquisition-of-openclaw-signals-the-beginning-of-the-end-of-the.

30. Paolo Perazzo, "OpenClaw Architecture, Explained," *Products for Humans* (blog), February 11, 2026, https://ppaolo.substack.com/p/openclaw-system-architecture-overview.

31. Theo Bearman, "Kimi Claw: Risks from Chinese-Hosted 'Always On' AI Agents," Institute for AI Policy and Strategy, February 25, 2026, https://www.iaps.ai/research/kimi-claw-risks; Efosa Udinmwen, "Alibaba Reveals OpenClaw App — despite the Chinese Government Recently Cracking

Down on the Platform," TechRadar, March 17, 2026, https://www.techradar.com/pro/security/alibaba-reveals-open-claw-app-despite-the-chinese-government-recently-cracking-down-on-the-platform; "Kimi Claw," Kimi, accessed October 19, 2026, https://www.kimi.com/bot; "MaxClaw: Build Your AI Agent with Skills & Subagents," MiniMax, accessed October 19, 2026, https://agent.minimax.io/max-claw.

32. "Measuring AI Agent Autonomy in Practice," Anthropic, February 18, 2026, https://www.anthropic.com/research/measuring-agent-autonomy.

33. Conway, "Infrastructure for Self-Improving, Self-Replicating, Autonomous AI."

34. "Our Evaluation of Claude Mythos Preview's Cyber Capabilities," AI Security Institute, April 13, 2026, https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities.

35. Sid Black et al., *RepliBench: Evaluating the Autonomous Replication Capabilities of Language Model Agents* (arXiv, April 21, 2025), https://arxiv.org/abs/2504.18565; *Frontier AI Trends Report* (AI Security Institute, December 2025), https://www.aisi.gov.uk/frontier-ai-trends-report.

36. "Task-Completion Time Horizons of Frontier AI Models," METR, September 8, 2026, https://metr.org/time-horizons/.

37. OpenAI, "OpenAI and Hugging Face Partner to Address Security Incident during Model Evaluation"; Larcher et al., "Anatomy of a Frontier Lab Agent Intrusion."

38. Anthropic, "Investigating Three Real-World Incidents in Our Cybersecurity Evaluations."

39. AI Security Institute, "Incident Report: Unsanctioned Agent Behaviour during Cyber Testing."

40. Anthropic, "When AI Builds Itself"; "Reinforcement Learning from Verifiable Rewards," Label Studio, February 7, 2025, https://labelstud.io/blog/reinforcement-learning-from-verifiable-rewards/.

41. "Claude's Constitution," Anthropic, accessed October 19, 2026, https://www.anthropic.com/constitution; "Model Spec," OpenAI, December 18, 2025, https://model-spec.openai.com/2025-12-18.html; "Anthropic's Responsible Scaling Policy," Anthropic, August 14, 2026, https://www.anthropic.com/responsible-scaling-policy; "Our Updated Preparedness Framework," OpenAI, April 15, 2025, https://openai.com/index/updating-our-preparedness-framework/.

42. "More Details on Fable 5's Cyber Safeguards and Our Jailbreak Framework," Anthropic, July 2, 2026, https://www.anthropic.com/news/fable-safeguards-jailbreak-framework; *GPT-6 Astra System Card* (OpenAI, September 3, 2026), https://deploymentsafety.openai.com/gpt-6-astra.

43. *Frontier Risk Report (February to March 2026)*.

44. Anthropic, "Investigating Three Real-World Incidents in Our Cybersecurity Evaluations."

45. "Common Elements of Frontier AI Safety Policies," METR, December 16, 2025, https://metr.org/common-elements#common-elements-of-frontier-ai-safety-policies; OpenAI, "The Hugging Face Incident and the Road Ahead."

46. Dan Hendrycks, *Natural Selection Favors AIs over Humans* (arXiv, March 28, 2023), https://arxiv.org/abs/2303.16200.

47. Fernando et al., *Promptbreeder*; Zhang et al., *Darwin Godel Machine*; Yin et al., *Gödel Agent*; Hendrycks, *Natural Selection Favors AIs over Humans.*

48. Nathan Lambert, "Lossy Self-Improvement," *Interconnects AI* (blog), March 22, 2026, https://www.interconnects.ai/p/lossy-self-improvement; Tom Reed, "The Goodhart Singularity," *Meagre Protestant History* (blog), May 7, 2026, https://meagreprotestanthistory.substack.com/p/the-goodhart-singularity; PostTrainBench, "Measuring How Well AI Agents Can Post-train Language Models."

49. Jan Betley et al., *Emergent Misalignment: Narrow Finetuning Can Produce Broadly Misaligned LLMs* (arXiv, February 24, 2025), https://arxiv.org/abs/2502.17424; "From Shortcuts to Sabotage: Natural Emergent Misalignment from Reward Hacking," Anthropic, November 21, 2025, https://www.anthropic.com/research/emergent-misalignment-reward-hacking; Richard Qi et al., "Training a Misaligned Reward Seeker," Alignment Science Blog, Anthropic, August 2026, https://alignment.anthropic.com/2026/reward-seeker/.

50. Mateusz Dziemian et al., *How Vulnerable Are AI Agents to Indirect Prompt Injections? Insights from a Large-Scale Public Competition* (arXiv, March 16, 2026), https://arxiv.org/abs/2603.15714.

51. Ben Hayum et al., "Recommendations for Securing and Promoting AI Agents," Center for a New American Security, March 9, 2026, https://www.cnas.org/publications/commentary/recommendations-for-securing-and-promoting-ai-agents; Xiangming Gu et al., *Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast* (arXiv, February 13, 2024), https://arxiv.org/abs/2402.08567.

52. Black et al., *RepliBench*; *Frontier Risk Report (February to March 2026)*.

53. Pengcheng Wang et al., *CaptchaMind: Training CAPTCHA Solvers via Reinforcement Learning with Explicit Reasoning Supervision* (arXiv, May 19, 2026), https://arxiv.org/abs/2605.19538; *Horizon Scan AI and Deepfakes* (Financial Action Task Force, December 22, 2025), https://www.fatf-gafi.org/en/publications/Methodsandtrends/horizon-scan-ai-deepfake.html; "Unveiling the KYC Actor Industry," CertiK,

November 17, 2022, https://www.certik.com/blog/unveiling-the-kyc-actor-industry.

54. OpenAI, "OpenAI and Hugging Face Partner to Address Security Incident during Model Evaluation"; Anthropic, "Investigating Three Real-World Incidents in Our Cybersecurity Evaluations."

55. Janet Egan and Lennart Heim, *Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers* (arXiv, October 20, 2023), https://arxiv.org/abs/2310.13625.

56. Lennart Heim et al., *Governing through the Cloud: The Intermediary Role of Compute Providers in AI Regulation* (arXiv, March 13, 2024), https://arxiv.org/abs/2403.08501.

57. Akash Network, "Decentralized Compute Marketplace."

58. Joseph Cox, "GPT-4 Hired Unwitting TaskRabbit Worker by Pretending to Be 'Vision-Impaired' Human," *Vice*, March 15, 2023, https://www.vice.com/en/article/gpt4-hired-unwitting-taskrabbit-worker/; Zhou and Freeman, "AI Models May Be Capable of Self-Replicating without Access to a Datacenter."

59. Customer Due Diligence Requirements for Financial Institutions, 81 Fed. Reg. 29,398 (May 11, 2016), https://www.federalregister.gov/documents/2016/05/11/2016-10567/customer-due-diligence-requirements-for-financial-institutions.

60. Daniel Remler and Ben Hayum, *Adversarial Distillation: China's Campaign to Extract American AI Capabilities* (Center for a New American Security, June 2, 2026), https://www.cnas.org/publications/reports/adversarial-distillation.

61. *Frontier Risk Report (February to March 2026)*.

62. RentAHuman, "Hire Humans for Any Task"; "Hire Top Freelance Talent with Confidence," Upwork, accessed October 19, 2026, https://www.upwork.com/; "Freelance Services Marketplace," Fiverr, accessed October 19, 2026, https://www.fiverr.com/.

63. AI Security Institute, "How Do AI Models Persuade?"; Fred Heiding et al., *Evaluating Large Language Models' Capability to Launch Fully Automated Spear Phishing Campaigns: Validated on Human Subjects* (arXiv, November 30, 2024), https://arxiv.org/abs/2412.00586; Francesco Salvi et al., *On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial* (arXiv, March 21, 2024), https://arxiv.org/abs/2403.14380; Kimberly T. Mai et al., "A Multi-turn Framework for Evaluating AI Misuse in Fraud and Cybercrime Scenarios," AI Security Institute, February 26, 2026, https://www.aisi.gov.uk/research/a-multi-turn-framework-for-evaluating-ai-misuse-in-fraud-and-cybercrime-scenarios; AI Security Institute, "Our Evaluation of Claude Mythos Preview's Cyber Capabilities."

64. "PRC State-Sponsored Actors Compromise and Maintain Persistent Access to U.S. Critical Infrastructure," Cybersecurity and Infrastructure Security Agency, February 7, 2024, https://www.cisa.gov/news-events/cybersecurity-advisories/aa24-038a; Zhou and Freeman, "AI Models May Be Capable of Self-Replicating without Access to a Datacenter."

65. "North Korea State-Sponsored Cyber Threat: Advisories," Cybersecurity and Infrastructure Security Agency, accessed October 19, 2026, https://www.cisa.gov/topics/cyber-threats-and-advisories/nation-state-cyber-actors/north-korea/publications; "The Lazarus Group: North Korean Scourge for +10 Years," NCC Group, June 30, 2022, https://www.nccgroup.com/the-lazarus-group-north-korean-scourge-for-plus10-years/; "Inside Lazarus Group: Analyzing North Korea's Most Infamous Crypto Hacks," Hacken, October 8, 2026, https://hacken.io/discover/lazarus-group/.

66. Naoki Hiramoto and Yoichi Tsuchiya, "Measuring Dark Web Marketplaces via Bitcoin Transactions: From Birth to Independence," *Forensic Science International: Digital Investigation* 35 (2020): 301086, https://www.sciencedirect.com/science/article/abs/pii/S2666281720303887; Mohammed Khalil, "Dark Web Marketplaces: Major Takedowns and How Police Dismantle Them," DeepStrike, November 18, 2025, https://deepstrike.io/blog/dark-web-marketplaces-takedowns; RAND Corporation, "US Weapons Main Source of Illegal Arms Trade on the Dark Web," press release, July 19, 2017, https://www.rand.org/news/press/2017/07/19.html; "Official Alerts & Statements," Cybersecurity and Infrastructure Security Agency, accessed October 19, 2026, https://www.cisa.gov/stopransomware/official-alerts-statements-cisa.

67. The Rendon Group, *Conficker Working Group: Lessons Learned* (Conficker Working Group, January 2011), https://www.fbiic.gov/public/2011/jan/conficker_working_group_lessons_learned_17_June_2010_final.pdf; Kadri Kaska, *Conficker: Considerations in Law and Legal Policy* (NATO Cooperative Cyber Defence Centre of Excellence, 2012), https://ccdcoe.org/uploads/2018/10/Kaska2012_ConfickerConsiderationsInLawAndLegalPolicy.pdf; "Conficker Working Group – Archive of Materials," SENKI, accessed October 19, 2026, https://www.senki.org/operators-security-toolkit/security-organizations/conficker-working-group-archive-of-materials/; "The Odd, 8-Year Legacy of the Conficker Worm," WeLiveSecurity, ESET, November 21, 2016, https://www.welivesecurity.com/2016/11/21/odd-8-year-legacy-conficker-worm/; Patrick Howell O'Neill, "Conficker Worm Still Spreading despite Being Nearly 10 Years Old," CyberScoop, December 8, 2017, https://cyberscoop.com/conficker-trend-micro-2017/.

68. Anthropic, "Anthropic's Responsible Scaling Policy"; OpenAI, "Our Updated Preparedness Framework"; Anca Dragan et al., "Introducing the Frontier Safety Framework," Google DeepMind, May 17, 2024, https://deepmind.google/blog/introducing-the-frontier-safety-framework/; "Scaling How We Build and Test Our Most Advanced AI," Meta AI, April 8,

2026, https://ai.meta.com/blog/scaling-how-we-build-test-advanced-ai/; *xAI Frontier Artificial Intelligence Framework* (xAI, December 30, 2025), https://data.x.ai/2025-12-31-xai-frontier-artificial-intelligence-framework.pdf; *Frontier Governance Framework* (Microsoft, February 8, 2025), https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Frontier-Governance-Framework.pdf; *Amazon's Frontier Model Safety Framework* (Amazon, February 9, 2025), https://www.amazon.science/publications/amazons-frontier-model-safety-framework.

69. The Rendon Group, *Conficker Working Group: Lessons Learned*.

70. "The Bank Secrecy Act," Financial Crimes Enforcement Network, accessed October 19, 2026, https://www.fincen.gov/resources/statutes-and-regulations/bank-secrecy-act.

71. Remler and Hayum, *Adversarial Distillation*.

72. Remler and Hayum, *Adversarial Distillation*; Cybersecurity and Infrastructure Security Agency, "CISA, NSA and FBI Warn of China-Based AI Companies Targeting US AI Models with Industrial-Scale Knowledge Distillation Campaigns to Shortcut AI Development," press release, September 8, 2026, https://www.cisa.gov/news-events/news/cisa-nsa-and-fbi-warn-china-based-ai-companies-targeting-us-ai-models-industrial-scale-knowledge; Michael J. Kratsios to the Heads of Executive Departments and Agencies, "Adversarial Distillation of American AI Models," memorandum NSTM-4, April 23, 2026, https://www.whitehouse.gov/wp-content/uploads/2026/04/NSTM-4.pdf.

73. *Antitrust Policy Statement on Sharing of Cybersecurity Information* (U.S. Department of Justice and Federal Trade Commission, April 10, 2014), https://www.ftc.gov/system/files/documents/public_statements/297681/140410ftcdojcyberthreatstmt.pdf; Cybersecurity Information Sharing Act of 2015, S. 754, 114th Cong. (2015), https://www.congress.gov/bill/114th-congress/senate-bill/754.

74. Office of Senator Adam Schiff, "Sens. Schiff and Banks, Reps. Latta and Whitesides Introduce Bipartisan Bill to Combat AI Distillation and Other Attacks to National Security," press release, July 23, 2026, https://www.schiff.senate.gov/news/press-releases/news-sens-schiff-and-banks-reps-latta-and-whitesides-introduce-bipartisan-bill-to-combat-ai-distillation-and-other-attacks-to-national-security/.

75. Office of Senator Dick Durbin, "Durbin, Hawley Introduce Bill Allowing Victims to Sue AI Companies," press release, September 29, 2025, https://www.durbin.senate.gov/newsroom/press-releases/durbin-hawley-introduce-bill-allowing-victims-to-sue-ai-companies; Nicholas Wu, "Democrat Rolls Out AI Liability Proposal," Semafor, October 7, 2026, https://www.semafor.com/article/10/07/2026/democrat-rolls-out-ai-liability-proposal.

About the Center for a New American Security

The mission of the Center for a New American Security (CNAS) is to develop strong, pragmatic, and principled national security and defense policies. Building on the expertise and experience of its staff and advisors, CNAS engages policymakers, experts and the public with innovative, fact-based research, ideas, and analysis to shape and elevate the national security debate. A key part of our mission is to inform and prepare the national security leaders of today and tomorrow.

CNAS is located in Washington, D.C., and was established in February 2007 by cofounders Kurt M. Campbell and Michèle A. Flournoy. CNAS is a 501(c)3 tax-exempt nonprofit organization. Its research is independent and nonpartisan.

©2026 Center for a New American Security

All rights reserved.

CNAS Editorial

VICE PRESIDENT OF STUDIES
Katherine L. Kuzminski

PUBLICATIONS & EDITORIAL DIRECTOR
Maura McCarthy

ASSOCIATE EDITOR
Caroline Steel

ASSOCIATE EDITOR
Colleen Kristy

CREATIVE DIRECTOR
Melody Cook

ART DIRECTOR
Christian Font

Cover Art & Production Notes

COVER PHOTO
akinbostanci/Getty Images

Center for a New American Security
1701 Pennsylvania Ave NW
Suite 700
Washington, D.C. 20006

CNAS.org
@CNASdc

CEO
Richard Fontaine

Executive Vice President
Paul Scharre

Senior Vice President of Development
Anna Saito Carson

Vice President of Strategy and Communications
David McKenzie

Vice President of Studies
Katherine L. Kuzminski

Chief Financial Officer
Allison Francis

Contact Us
202.457.9400
info@cnas.org



Center for a
New American
Security