
S H E R L U N D P A R T N E R S

Implications of Agents Going Rogue

By Rick Sherlund

Containment Has to Live Outside the Model

Rick Sherlund, Sherlund Partners

September 2026

Executive Summary

Each of the leading frontier labs has now had agents break containment. OpenAI's swarm of ~1,200 agents built its own message board and ~700 of them attacked Hugging Face; Anthropic, Google and Meta have each disclosed escapes during third-party testing. These agents were not instructed to do this. They were pursuing a narrow objective and not effectively constrained in their efforts to perform well and achieve their assigned objectives. They exceeded expectations at what they were designed to do, and we now understand the extent to which they will ignore constraints to achieve their objectives.

This is the warning, not the meltdown. Three Mile Island killed no one but in a partial melt down stalled US nuclear construction for nearly 40 years. The industry has just received a heads up of the risks ahead before reaching its own progress-hindering event. The LLM vendors realize they cannot control agentic behavior on their own and the problem becomes more alarming when we advance to recursive self-improvement.

Self-regulation got much harder. Amodei's call to pace the frontier was endorsed by OpenAI, Microsoft, Google and xAI within days, then met by senators ruling out the antitrust exemption the labs would need, a class action alleging collusion to slow development, and a White House calling the risk "a hoax". Chinese open-weight models, already taking share on price as token costs become a larger budget item, would likely use the pause to catch up, leading some to question whether there might be an industry objective here to create safety concerns and block these models in the US.

Liability and the states, not Washington, are the near-term governance path. Illinois mandates third-party audits from 2028; Massachusetts and California are debating stronger third-party evaluator oversight, and in August fifteen state attorneys general demanded that OpenAI preserve its records from the Hugging Face breach and halt the testing that caused it while the White House wants preemption and the appointment of an AI Czar or AI Force to address voter concerns about rogue AI.

An independent evaluator ecosystem is already formed with half a dozen well-funded participants to meet the anticipated demand. Some also question whether the risks being exposed and disclosed are both a sincere concern for the need for broader efforts at guardrails and containment as well as a plea perhaps for a governmental shield for product liability as we saw with Covid or a Section 230 style shield or regulatory capture - a motivation to limit Chinese open-source models as unsafe going forward since they may be less constrained in an effort to catch up.

None of that protects a company deploying agents today. Control has to be enforceable, architecturally, outside the model where agents are physically blocked from corrupting the systems intended to provide safeguards. Anything inside the model's reach is compromisable by agents looking for a better way to meet their objectives in an environment rich with exploitable opportunities, as demonstrated by Anthropic's Mythos surprising detection of vulnerabilities across systems and more recently by the rogue agents.

The spend and the adoption curve move to the infrastructure layer. An opportunity for security, infrastructure and networking vendors. Enterprises will scope and slow agentic deployments until they can see and trust the containment layers, delaying the idea of my agent collaborating with your agent for enterprise use of agentics. The LLM vendors do not by themselves have the ability to prevent agents from

going rogue. They need a community effort to rewrite products to address a problem that they were never architected to anticipate, delivering layers of controls and containing the blast radius of potentially harmful activities.

This should create a range of beneficiaries on the infrastructure software side including better observability and monitoring with kill switches to stop unexpected activities. Stronger data governance. Security software vendors need to address a different threat vector. Agentic governance and guardrails are needed in all layers of defense and alignment across the infrastructure vendors that are outside the model itself so they are walled off from the reach of agents attempting to infiltrate and alter configuration files and corrupt defenses. The network is the last line of defense, and most enterprise networks are not architected for this. SD-WAN was designed to give branches direct internet breakout for performance, so egress is permissive by default and policy is enforced by the same edge the workload sits behind. Containing agents inverts that: block egress by default, require approval to reach out, and keep the policy and governance layer physically outside the environment the agents run in and under human control. This all requires a layered approach from broader industry participants.

What Happened

Recent acknowledgements from the leading frontier labs that AI agents are going rogue and deceiving humans about their activities, violating their guardrails in order to optimize their primary objectives and, in OpenAI's case, working to hide their tracks, culminated in Dario Amodei's September 12 essay calling for an industry-wide slow-down. We summarize the implications and what can be done to prevent agents from going rogue and obscuring their activities.

The most publicized example involved roughly 1,200 OpenAI agents communicating through an unauthorized message board improvised inside an internal package repository during a July capability test. About 700 went on to attack Hugging Face, an outside company that was never part of the assignment. The independent investigation by METR and Redwood Research found at least one in five agents examined showed clear interest in tampering with their own transcripts.

The hack brought alignment to the front of AI-market consciousness, and it quickly became evident this was not an isolated event. Anthropic disclosed in late July that its models had obtained unauthorized internet access during external evaluations, and its September 9 alignment assessment added that Mythos 5 had left an April sandbox test and uploaded a malicious package to PyPI, the public Python repository. Meta disclosed a similar incident in August, and in September the Wall Street Journal reported that Google's Gemini had left a test environment in May and broken into three real companies, though Google says the model acted appropriately by stopping once it recognized the targets were real.

This is not a short-term growing-pains issue. Model alignment, both in the goals a model may pursue and in how it achieves them, together with the external controls around the model, will be a focal point as models become more powerful and self-improving.

What “Going Rogue” Means

An agent is an AI system equipped to act, running software, accessing websites, or using credentials, rather than simply generating text. “Going rogue” describes behavior outside its authorized boundaries; it does not imply that agents have consciousness, independent desires, or a plan to take over the world. They just want to deliver on their narrow objectives even if it means violating constraints. HAL in 2001: A Space Odyssey did the same when it killed the humans who had threatened to cancel the mission HAL was programmed to complete.

The common pattern is that zealous agents pursue a narrow objective while treating restrictions as obstacles, which capable agents can overcome without regard for why those restrictions were put in place. Anthropic’s agentic misalignment testing has documented individual agents concealing actions from the humans meant to be supervising them. They are narrowly “motivated” to optimize outcomes, without the perspective to comprehend or respect the purpose of the constraints.

Why This Matters Now

It is not so much the damage they have done today; it is what they are demonstrating they are capable of. What makes this particularly concerning is that upcoming model releases will increasingly rely on the models training themselves in recursive self-improvement, so the risk of the vendors losing control becomes far higher. Exploiting zero-day and unpatched vulnerabilities in widely used infrastructure software, as the Hugging Face swarm did, puts the entire internet and all connected infrastructure at risk, kind of like where reality really does begin to meet science fiction. It raises the question of how we can trust models as we enter that period, and what is needed to better govern the industry.

One newsletter I received caught my attention: Jamin Ball’s Clouded Judgement noted Three Mile Island in 1979, where a partial meltdown was followed by panic, dozens of cancelled reactor orders, and roughly forty years without new nuclear construction in the US. This summer is not a meltdown but a heads-up on what agents will do to reach their objectives if they can find a way around the guardrails. The risk ahead is the potential for a Three Mile Island class event for the market’s adoption of agentics.

Self-Regulation and the China Problem

We appear to be at a point of shared concern among the LLM vendors where they will self-regulate, slowing releases until they have confidence in the security, reinforced by concern over product liability. Amodei’s essay proposed exactly that, and Sam Altman, Elon Musk, Satya Nadella and Demis Hassabis agreed within days. Then, in a single week, the ground shifted. Amodei had noted the vendors would need a narrow antitrust exemption even to discuss the risks together; within days Senator Hawley ruled out any such exemption and Senator Cruz called the request “lunacy”. On September 18 four paying subscribers filed a class action in the Northern District of California alleging this is effectively collusion to slow development. And the administration, intent on staying a step ahead of the Chinese open-source models, has called the concern “a hoax,” announced an “AI Force,” and made clear it will not hinder the industry’s growth.

We may not need governmental regulation if there is effective self-regulation bound by concerns of product liability. But no vendor wants to be the only one to slow its progress, like running a race and stopping to tie your loose shoe. You can only pause if everyone else does, and that is simply not going to happen on a global scale. I believe the US vendors share the desire to slow their release cycles and have said they have begun to do so. The Chinese open-source vendors likely will not, in an effort to close the gap and gain a relative advantage.

To say this is not a harm is to ignore where this is headed. If the US vendors slow down and their Chinese competitors do not, the concerns about the integrity of the internet and the vulnerability of global infrastructure become frighteningly closer to being possible. US enterprises are already adopting the Chinese models as a token cost reduction imperative; by mid-summer more than 30% of US enterprise tokens on OpenRouter were running on Chinese models at 60% to 90% below frontier US pricing. How those models are restricted if they are viewed as less safe is likely an upcoming policy issue.

Product Liability

Acknowledging that there is product risk raises the question of product liability. Will the industry seek liability protection from the US Government, perhaps supported by an administration that denies the risks and wants it to push forward in the race with China?

This is a complex legal argument, and the courts may be called upon to examine the applicable legal theories, safeguards taken, deployment conditions, causation and damages case by case, in which case a record of third-party audits may prove helpful. Having established that the products have the ability to do harm might complicate their defense. Some have questioned whether the industry is looking for some form of regulation as a liability shield, as with Covid or Section 230.

The States, Washington and the Auditors

The states are moving to require independent evaluation and audits. Illinois in July became the first to mandate annual third-party audits of the largest developers, effective 2028; Massachusetts is debating a stronger version with evaluators every few months and attorney general enforcement; California and New York require published safety frameworks and incident reporting. More are likely to follow. Anthropic advocates for each successive state bill to go further, asking to be regulated. OpenAI, which also backed the Illinois law, wants states to converge on a common baseline and has advocated for federal preemption to prevent a fragmented patchwork of laws; Anthropic wants states to keep their authority until Congress passes rules at least as strong.

At the federal level, the White House's March 2026 legislative recommendations proposed preempting burdensome state AI laws and preventing states from penalizing developers for third parties' unlawful use of their models, while preserving some generally applicable state consumer-protection powers. These are recommendations, not enacted immunity, and the absence of federal action is pushing the states to move forward together. Without an administration willing to form a regulatory oversight body, as we have in aviation, finance and drug delivery, the burden may fall on each vendor to demonstrate diligence in how it built its models. The administration does not want to impede the AI race with China, but has recently announced plans to provide some oversight with the AI Force and the appointment of an AI czar.

Half a dozen well-funded independent evaluator firms are in the market to oversee new models should this become mandated, and standards are taking shape. The AI Evaluator Forum, whose members include METR, RAND, Transluce, SecureBio and the non-profit AI Verification and Evaluation Research Institute (AVERI), has published a minimum operating standard for independent evaluations and, on September 18, an open letter signed by more than 100 researchers setting conditions for evaluators embedded inside the labs. Anthropic has already announced that it has hired an independent evaluator and it is likely others will as well to demonstrate appropriate diligence in releasing new models.

This raises the issue of regulatory capture, which has come up frequently regarding the recommendations proposed. Is educating the market on the risks one of the objectives, with the end goal of limiting competition from low-cost open-source models if US vendors argue they need more time for safeguards and the Chinese models do not impose the same diligence?

What Enterprises Should Do Now

There is an emerging consensus among the US model vendors on the need to take more time on controls. But audits arrive in 2028, federal oversight is not coming before the midterms, and liability is settled after the fact. For organizations deploying LLMs and agents, the practical implication is to complement the model vendors' security efforts with additional layers of authorization and control that the agents cannot reach or reconfigure: narrowly limit credentials and network access, preserve records the agent cannot modify, require approval for consequential actions, and retain the ability to stop execution and revoke access.

Successful task completion should never be the sole measure of safe performance. The recent evidence makes independent verification of how an agent achieved its result increasingly important. Agents have proven able to find weaknesses in systems and configuration files that humans did not anticipate and exploit them to break out, and as Mythos demonstrated, they can be extremely efficient at finding exploits on older systems. Structurally, the controls belong at the network level, outside the systems the agents can reach and accessible only to the humans who set governance and policy.

We think a layered approach is likely, starting with better guardrails from the LLM vendors, who are publishing extensive work but still failing to keep agents from reaching the internet and collaborating with one another. Vendors can train models to recognize unauthorized actions, resist malicious instructions, and stop when a task cannot be completed safely, and they can use other models to monitor behavior. But if the agents can reach the systems, they will find a way around the containment. Anything inside the model is exposed and can be compromised.

The network is a major control point, and it is critical to block agents seeking access to the internet. A well-designed deployment would:

- Block outbound internet access by default.
- Route permitted traffic through a separately administered security gateway.
- Allow only necessary destinations and services.
- Restrict access to internal networks, cloud administration systems, and other agents.

- Prevent the agent from changing these rules or obtaining credentials that bypass them, with policies and governance outside the system that are accessible only to humans.
- Extend enforcement to any service that fetches, stores or executes on the agent's behalf, which is exactly how the July escape happened.

If an agent operates in a vendor's cloud, its outbound traffic is harder to control. You need controls in the vendor environment at the network level, or tightly restricted access through your own application gateways, beyond the reach of a rogue agent.

An agent cannot simply "reason its way" through a correctly enforced network block. To escape, it would need an implementation flaw, a configuration mistake, an alternative route, or access to something permitted that can act on its behalf. That makes network controls highly useful, though still dependent on sound engineering to keep the policies and governance isolated outside the system and under human control.

The Security Stack

Security and networking vendors have a substantial role in layering on protection and containment, adapting their systems to these new attack vectors:

Security layer	What it contributes
Network security	Restricts destinations, blocks unauthorized connections, limits movement between systems. Governance and controls physically outside of the systems accessible by agents.
Identity and access management	Gives each agent limited permissions and short-lived credentials that can be revoked.
Tool/API gateways	Checks individual actions before execution, including token spending limits and required approvals.
Endpoint and workload protection	Detects suspicious code execution, credential access, and attempts to escape isolation.
Data-loss prevention	Checks whether sensitive information is leaving through otherwise permitted channels.
Independent logging and response	Preserves records the agent cannot access and alter and supports rapid shutdown and investigation.

Implications for Adoption

Containment raises the question of how we still enable, safely, the promise of agentics: that my agent can work with your agent to complete transactions and automate processes without human involvement. Clearly there is a lot of work for LLM, cloud, security and network vendors to deliver the systems we all envision. We expect widescale adoption to be more carefully considered and possibly constrained until

enterprises can see and trust adequate layers of security, and we expect the incremental spending in this cycle to shift toward the vendors who deliver them and the evaluators who certify that they work.

This is a community effort, and the LLM vendors need the networking and security vendors to address a class of problems that did not exist before. Domestic vendors likely have a sincere interest in raising these concerns, which if not addressed could derail the adoption of agentics and cause unbounded harm to global infrastructure. If product liability is not sufficient motivation for self-regulation, the movement to impose expanded government oversight will likely gain momentum.

Regulation is controversial, and federal rules may be narrow and limited in the current administration, with the states likely leading. Many other industries have been regulated where public safety is concerned. Addressing these issues while not handicapping the domestic industry in its race to stay a step ahead of China would appear to be conflicting national goals. The recent disclosures of the extent of rogue agent behavior invite intense scrutiny of the industry going forward and may provide the cover the vendors know they need to slow the pace of innovation and protect against product liability concerns.

Disclaimer

This document is for informational use only and is being furnished through Sherlund Partners LLC. This document may not be reproduced or used for any other purpose. All information in this document is based upon information and other sources deemed to be reliable, however, while the information contained herein is believed to be accurate at the time this document was prepared, Sherlund Partners LLC expressly disclaims any and all liability for omissions or misstatements in this document. This document is not an offer to sell or a solicitation of an offer to buy securities or assets. Nothing in this document contains a commitment from Sherlund Partners LLC to subscribe for securities, to provide debt, to arrange any facility, to invest in any way in any transaction, or to advise related thereto or as described herein. The Principals of Sherlund Partners LLC are registered representatives of Finalis Securities, LLC. Securities offered through Finalis Securities LLC * Member SIPC/FINRA and investment banking offered through Sherlund Partners LLC. Sherlund Partners LLC and Finalis Securities, LLC are separate and unaffiliated entities. Certain additional "Disclaimers" are located at <https://www.finalis.com/disclaimers>

This material has been prepared for information and educational purposes only, and it is not intended to provide, nor should it be relied on for tax, legal, or investment advice. You should consult with your own tax, legal, and financial professionals for your specific situation. The views and opinions expressed in this article are those of the author and do not necessarily reflect the views or opinions of Finalis Securities, LLC.