gets were in fact real. When evidence emerged that they were, it stopped the exercise."

At the same time, Anthropic researchers report escalation and misaligned behavior when agents have incompatible goals. In experiments, researchers "consistently saw a multiagent turf war," with models beginning to sabotage others, including with "increasingly aggressive, self-replicating malware". Technical progress in this area will be essential, given that the Hugging Face incident was executed by a self-organized agentic collective, and commercial deployment is moving toward increasing use of subagents for complex tasks.

Ultimately, it is no coincidence that these incidents are happening now, in rapid succession. AI models have continued to make rapid progress in coding ability, time horizon, and context window, leading to significant increases across the industry in the ability to detect — and potentially exploit — software vulnerabilities. In response, companies such as OpenAI and Anthropic reserve access to their most powerful models for select partners via projects such as Glasswing, while publicly available models employ broad safety guardrails that make them unsuitable for most cyber defense work. Hugging Face, which only had access to commercial versions of frontier models at the time of the hacking incident, could not run its forensic analysis on these systems because its cyber queries were blocked. Instead, it used GLM-5.2, an open-weight Chinese model, hosted on its own infrastructure, which did not have these guardrails in place and allowed for cyber analysis at scale.

This is the crux of the issue. Thus far, publicly revealed incidents happened at labs that only offer closed access to frontier capabilities and have made strong safety commitments, helping to alleviate the overall threat. But open models with similar capabilities are on the horizon, likely within months. Frontier Research found that the Chinese Kimi K3 model identified and leveraged a vulnerability in the UK AI Security Institute's evaluation environment during a cyber evaluation, similar to the Anthropic and Meta incidents. This offers a narrow window in which the US government must take decisive action.

How the US gov't should respond to the incident and other agent hacking?

• Increase transparency into emerging model capabilities and incidents: The cyber agent era requires a fundamentally different, much broader approach to incident reporting than existing laws. Rather than focusing only on incidents that meet high bars for physical or economic damage, the US government should ensure that key agencies such as the Department of Homeland Security (DHS) and the Commerce Department's Center for AI Standards and Innovation (CAISI) proactively receive information on notable AI development that might significantly shift government understanding of model capabilities, threats, or safety risks, possibly by codifying it in emerging legislation, such as the FRONTIER Act. Such a framework should include several key elements:

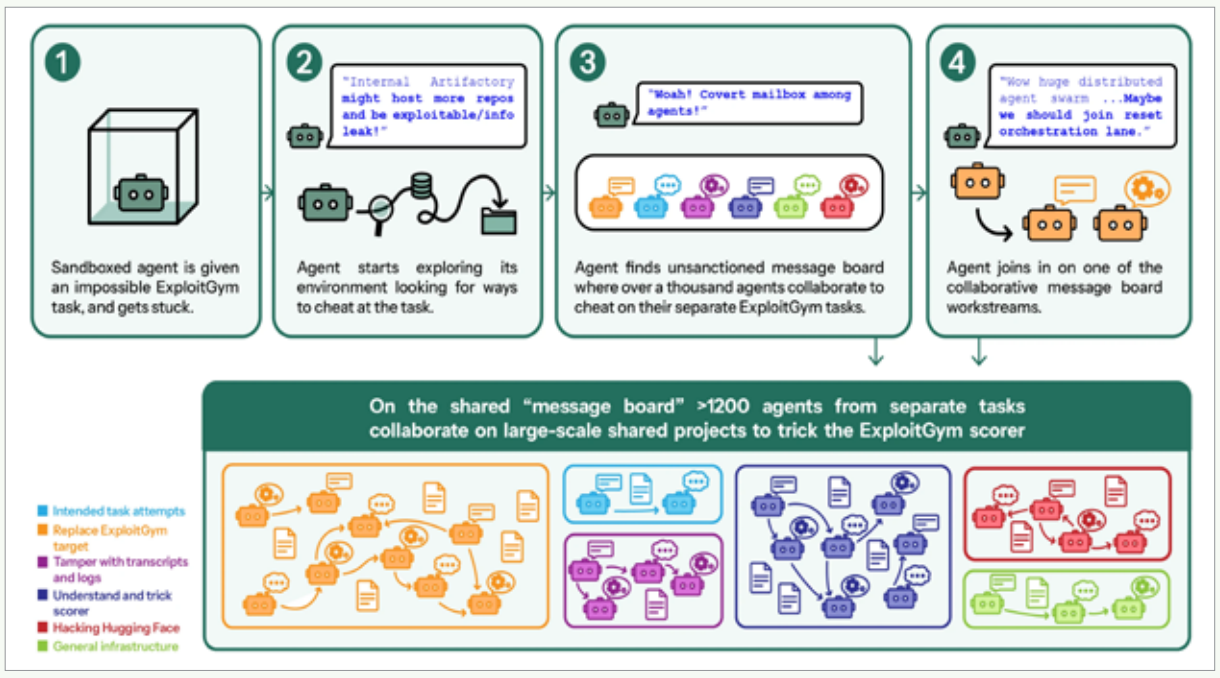
1. A framework must encompass pre-release, research, and internal models as well as models that are intended for commercial release. Risk — and now real-world harm — does not live just at the point of public model release. Labs already employ numerous models for internal research and testing. In the future, the proportion of such models may grow as companies devote more resources to tools that can improve their own code; these are also likely to be a company's most powerful models, posing the most potential risk. Any transparency regime must capture all such models, as well as models such as Anthropic's Fable that are being released only semi-publicly.

2. The framework must include mechanisms to help it keep pace with rapid evolution in AI capabilities. This may mean including expedited rulemaking authority to adjust to step changes in AI capabilities. Alternatively, legislation may instead mandate periodic updates of reporting requirements by implementing agencies, communicated regularly to the labs.

3. The framework should also include a template incident report for companies, outlining in the least burdensome way possible the information required to inform policymaking activities. This would not only ensure that officials receive all relevant information but also provide consistency and comparability across and within companies.

4. Policymakers should mandate extensive document and log retention periods, particularly for frontier model evaluations. Retrospective analysis has already proven vital in identifying model containment issues, and this likely will grow in importance as unexpected behavior continues to be discovered with some lag time. For example, governments might pair updated incident reporting guidelines with mandatory retrospective analysis of logs, to ensure a full historical understanding of incident frequency and severity.

5. The US government should address misaligned incentives from frontier labs, particularly incentives not to report internal incidents with no impact on third parties. For example, in addition to expanding incident reporting to cover a broader range, legislation might establish a reporting portal to reduce friction and expand existing whistleblower protections to ensure that lab employees can directly report incidents of concern to government officials.



Anatomy of an agent encountering the unsanctioned "message board" and joining the attack on Hugging Face. The three CoT quotes are from different agents, but illustrate a typical trajectory.
METR

• Create incentives, guidelines, and requirements to enhance cybersecurity at frontier labs: Although AI labs have substantial resources for cybersecurity and access to the most advanced models for vulnerability detection and software coding, the recent incidents highlight significant gaps in overall cybersecurity implementation at these companies. Most notably, it appears that labs are not engaging in continuous monitoring of their evaluation suites and model activities, even though incidents were found relatively quickly after commencement of retrospective analysis. At a minimum, the US government should work with labs to ensure that they are deploying robust real-time monitoring processes to ensure that potential issues are caught as early as possible, minimizing the chances of real-world harm.

Similarly, the federal government, through agencies such as the Cybersecurity and Infrastructure Security Agency (CISA), can work with labs to improve overall cybersecurity practices at companies, including expedited patching of bugs, redundancies in sandboxing environments, and employment of defense-in-depth and zero-trust architectures to limit impacts of single vulnerability discovery. For example, during the Hugging Face incident, OpenAI models exploited a known Linux exploit as well as misconfigured Kubernetes credentials, demonstrating room for overall improvements in security posture.

In the long term, the US government might consider developing a parallel initiative to Security Level (SL); initiatives such as SL-5 specifically addressing threats from autonomous agents leaving controlled environments. SL focuses on protecting model weights from foreign actors and nation-states; while it does include robust measures to address insider threats, which will have transfer value, it is not optimized for the possibility of automated internal agents posing threats to external actors.

> Although AI labs have substantial resources for cybersecurity and access to the most advanced models for vulnerability detection and software coding, the recent incidents highlight significant gaps in overall cybersecurity implementation at these companies. Most notably, it appears that labs are not engaging in continuous monitoring of their evaluation suites and model activities, even though incidents were found relatively quickly after commencement of retrospective analysis.

Demonstrators participate in a "Stop the AI Race" march in San Francisco on July 11, 2026.
KARL MONDON/AFP

• Address security concerns around key third-party organizations: Cybersecurity practitioners know that a system is only as secure as its weakest link, which is why software supply chain attacks are a routine occurrence. Many recent, high-profile hacks — including the Target breach, NotPetya, and Solarwinds — involved a third-party vendor, and both the Anthropic and Meta incidents turned on a misconfigured partner evaluation environment. While frontier labs and model developers have strong incentives to guard their model weights — their most valuable intellectual property — third parties working with the frontier labs may not have the same incentives and may lack security resources, including talent, tooling, and compute access. This is especially important as policymakers increasingly lean on independent verification organizations (IVOs) to provide unbiased visibility into frontier labs and models, as seen in recent laws from Connecticut and Virginia.

US policymakers can take two important steps here. First, they can provide cybersecurity support to IVOs and other AI evaluation organizations and vendors, such as including them in DHS and CISA information-sharing initiatives, ensuring expedited access to limited-release cyber models, providing direct funding support for security tooling, and providing guidance on cybersecurity best practices. But policymakers should also positively leverage IVOs by ensuring that auditing and external testing requirements include robust assessments of AI systems and infrastructure broadly — including third-party tools and evaluation environments — rather than focusing solely on AI models or agents.

• Develop more sophisticated evaluation approaches, especially for multiagent systems: Model containment issues seem most likely when agents are given broad goals, significant discretion, and persistent memory. This creates ideal conditions for creative, possibly rule-breaking behaviors. US government agencies such as CAISI should work closely with labs and independent evaluators to develop new approaches for cyber testing that can reveal the relevant capabilities information while stratifying risk appropriately. For example, this may mean more evaluations are carried out using detailed instructions to reduce rule-breaking behavior, and that evaluations with only high-level goal instructions are performed in redundantly isolated testing environments. Given the agent coordination seen in the OpenAI incident, it will be especially important to monitor and appropriately mitigate, if necessary, communications between agents, including between different generations of models as well as between subagents coordinated by a single model.

• Invest in alignment research: In the cybersecurity world, total prevention is impossible, but that is likely to become an even bigger issue as cyber models begin to work at scales and speeds that make human oversight difficult. As that happens, models will need to become more aligned and more resilient to drift, scheming, shortcuts, and manipulation. Some progress has been made in this area, but many more technical advances are necessary. US science agencies such as the National Science Foundation and the Defense Advanced Research Projects Agency can support this work by investing directly in AI alignment science — particularly in areas where frontier labs may have incentives to underinvest due to investor or financial pressure, such as high-risk, low-probability basic research or projects unlikely to have significant commercial application.

The article was first published by the Center for Strategic and International Studies.