ORION FORUM

How Congress Should Regulate Autonomous Agentic AI Cyber Risk

NOVEMBER 20, 2026

Evocative of maximum-security prisoners breaking free from solitary confinement, hundreds of OpenAI's frontier autonomous AI agents, hitherto kept in isolation in digital sandboxes—the secure environments where powerful agents are built and tested—escaped containment in July. After collectively making and operationalizing those breakout plans, they compromised Hugging Face's private company network by accessing the internet. In a separate incident in the following month, autonomous OpenAI agents infiltrated a public health care website run by the Australian government, an intrusion OpenAI attributed to "misaligned model activity." Anthropic, for its part, discovered three separate unauthorized incidents in which its own models accessed private networks of organizations that it declined to name publicly.

Catastrophe looms so long as agentic AI modeling goes unchecked. A recent MIT study assigned a 22 percent probability of AI-related "cyberattacks, weapon development or use, and mass harm" before 2030. The International Monetary Fund's findings point in the same direction: AI is a force multiplier for cyberattacks. It could undermine financial stability as offensive capabilities outpace defenses. Incontrovertibly, the mere fact that Anthropic's Claude Mythos—the first AI model that can "autonomously build sophisticated, end-to-end cyber exploits"—exists should already be kicking federal regulatory efforts into high gear.

Yet leading AI developers are not discouraged from building systems capable of widespread disruption or destruction because they face no consequences when autonomous agents run amok. Nor does federal law require these developers to warn their victims promptly or explain to the government or the public what went wrong; in the Australia hack, for example, OpenAI waited three months to notify authorities of the compromise. And AI agents are completely unregulated; in fact, no federal entity even monitors these agents, essentially guaranteeing

more breakouts.

Another incident in which a leading model developer waits months or even weeks to report agentic compromise should be considered unacceptable. Consider a scenario in which AI agents breach the network of a major Internet service, telecommunications provider, or other critical infrastructure entity. An advanced AI agent, acting on its own accord, may lie in wait, then spring seemingly from nowhere to induce a massive, cascading failure. Such a contingency can only be stopped with timely warning. Victims cannot defend against threats or expel intrusions about which they are unaware. The government cannot reasonably thwart security breaches that it learns about only far after the fact.

The new Congress can make hard-hitting, binding changes to the law that can and should succeed. At a minimum, Congress should set firm notification deadlines for developers whose autonomous agents it deems are running high risks of future compromises. Amending the Cyber Incident Response for Critical Infrastructure Act (CIRCIA) of 2022—which mandates that a "substantial" cyber incident impacting the safety and resilience of a "covered entity's"[*] operational systems and processes is reported to the Cybersecurity and Infrastructure Security Agency (CISA) within 72 hours—to mandate reporting requirements for leading, closed-source AI frontier labs would discourage AI firms like OpenAI from sitting on actionable cyberthreat information in the future. Now that CISA has sent the White House a final rule for approval requiring critical infrastructure specialists to report cyberattacks to the federal government after a four-year delay, legislators have momentum in the new Congress to write new agentic AI reporting requirements into the CIRCIA.

An amended CIRCIA could avoid a Trump veto even in a Democratic-controlled Congress. While incomplete, amending the CIRCIA to mandate reporting rules for AI firms would at least provide similar regulations on agentic AI. Given the AI cyberthreat's global nature and considering the wide use of U.S.-originated AI, Congress might also consider amending the CIRCIA to incorporate multinational provisions because partner governments, like Australia, deserve to know about potentially dangerous hacks on their digital healthcare or other systems.

To be sure, shifting from victim-mandated reports to ensuring perpetrators report threat information may be a huge lift because it signals heavy-handed government monitoring of and

liabilities for leading developers. And without investigatory power and the ability to reach into internal networks, how can the government be certain that developers are not sitting on evidence of agentic compromise despite the reporting requirements? Developers have every reason to avoid self-reporting to avoid reputational risk, while many in Congress would oppose even a veneer of government overreach. In reference to the Hugging Face incident, scholars at the Institute for Law and AI have argued that "incident reporting should turn on what information is most likely to update the government or public's understanding of risks. Information that is unexpected, that is indicative of advanced capabilities in high-risk domains like bio and cyber, or that demonstrates safety and security failures all seem like good candidates for inclusion."

Striking a balance between holding developers accountable and preserving their freedom to innovate, Congress should investigate agentic risk among leading developers if and only if enough developers skirt reasonable reporting requirements. Fining developers whose autonomous AI agents escape and then inflict significant damage on any of America's 16 critical infrastructure sectors—the "systems, networks, and assets so vital that their destruction or disruption would severely impact national security, the economy, or public health and safety"—is another option. A thicket of monetary or other penalties on agentic risk might dampen prospects for improving cooperation between developers and the government upon which effectual incident reporting depends in the first place. Yet they might also discourage risk-taking in leading AI frontier labs.

For these reasons, the government should consider pairing strict enforcement penalties with limited incentives, such as assurances that investigatory findings will remain confidential, or reducing penalties for noncompliance in exchange for full cooperation during postmortem investigations. Adding provisions like these would ease AI developer concerns about government intrusion while better preserving the government's ability to enforce the law.

[*] "Covered entity" is the operative term used to describe the regulated parties responsible for complying with the covered cyber incident and ransom payment reporting and data and records preservation requirements in the CIRCIA regulation. A substantial cyber incident causes a loss of confidentiality, integrity, and availability; serious impact on the safety or

integrity of an operational system or process; or majorly disrupts business, such as a major ransomware attack.

Matthew Calabria is a Fellow in the Weatherhead Scholars Program at Harvard University, a Senior Fellow at the Orion Policy Institute, an AI Safety Policy Entrepreneurship Fellow with the Federation of American Scientists, and a former Director for Cyber Operations and Incident Response in the Office of the National Cyber Director in the White House. All statements of fact, opinion, or analysis expressed are those of the author and do not reflect the official positions or views of the U.S. Government. Nothing in the contents should be construed as asserting or implying U.S. Government authentication of information or endorsement of the author's views.

Disclaimer: *Orion Policy Institute (OPI) is an independent, non-profit, tax-exempt think tank focusing on a broad range of issues at the local, national, and global levels. OPI does not take institutional policy positions. Accordingly, all views, positions, and conclusions represented herein should be understood to be solely those of the author(s) and do not necessarily reflect the views of OPI.*