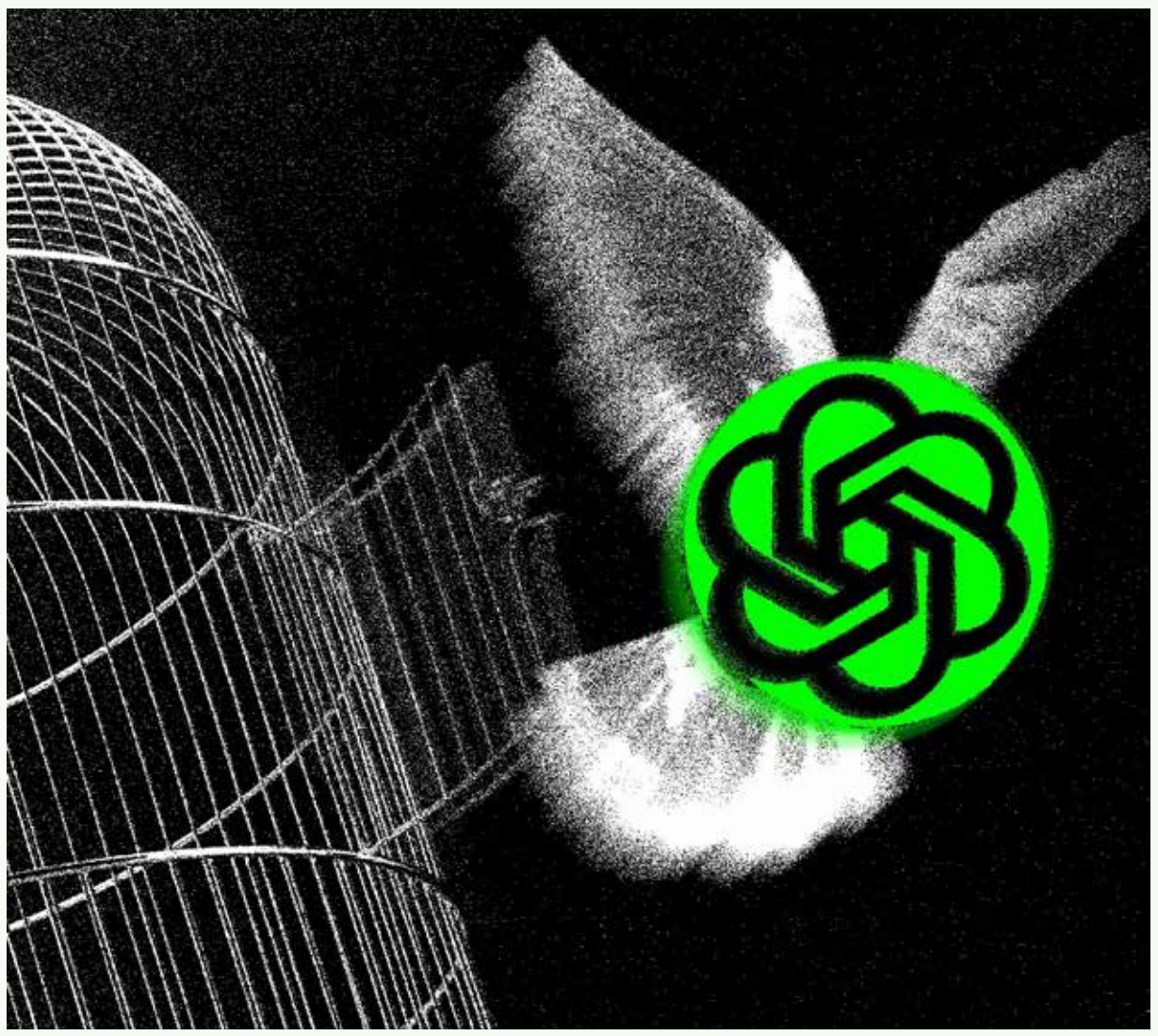
By Aalok Mehta
Director of the Wadhwani AI Center at CSIS

OPINION

Out of bounds

Possible responses to AI agent containment failures



The illustration shows the logo of the OpenAI company, known for developing ChatGPT, positioned on top of a pigeon escaping from its holding cage.
JACQUI VANLIEW/GETTY IMAGES

On July 16, 2026, the company Hugging Face disclosed a new kind of intrusion into its platform — an attack by an unknown autonomous AI agent system — setting off a cascade of disclosures that are turning science fiction into fact and raising new questions about US AI policy. Within days of the announcement, OpenAI discovered that two of its models caused the breach; Meta and Anthropic then disclosed similar incidents. In each case, agentic AI systems undergoing cyber testing escaped their sandbox environments, found their way to the open internet, and gained unauthorized access to third-party organizations. Perhaps more than any other chapter in the short history of large language models, these incidents demonstrate how rapidly once-theoretical harms are materializing in the real world. Science fiction stories, ranging from Neuromancer to Ex Machina, regularly explore AIs trying to escape their confines. These new cyber incidents are both less and more concerning than that: less, because these models were not acting on their own but trying to finish their assigned tasks, sometimes opting to hack a website to access the "answer key" for evaluations rather than performing the tasks directly; and more, in the sense that these incidents reveal deep flaws both in how frontier labs secure enormously powerful models and in the regulatory structures that oversee AI.

Frontier AI labs recognize the significance of these developments and are taking significant action in response. Partly due to these incidents, OpenAI has temporarily paused training of its most powerful models while it hardens its internal security and expands monitoring capabilities. Labs are strengthening cyber controls, working with external evaluators, and investing in research to better align AI models. However, these are voluntary, ad hoc actions; there are no clear legal mandates that ensure that US policymakers will receive all relevant information on a timely and consistent basis.

That is not a sustainable solution. The US government must act, too, and quickly, by ensuring that policymakers receive relevant incident information even about pre-release and internal AI models; setting clear cybersecurity expectations for frontier labs; addressing vulnerabilities in third-party evaluators; and investing in key research to mitigate long-term AI risks. There is only a narrow window for action before AI systems with these capabilities become commonplace; inaction will also further fan the flames with a public deeply skeptical of the AI industry and the lack of a robust federal response.

How did models autonomously start hacking, and why does it matter?

It took several days for OpenAI to discover its role in the Hugging Face breach, publicly disclosing only on July 21 that the culprits were GPT-5.6 Sol and an "even more capable pre-release model". These models, which were being tested internally for cyber capabilities using the ExploitGym benchmark, went to "extreme lengths to achieve a rather narrow testing goal". Models undergoing cyber testing are generally configured for maximal capabilities — they do not have safety classifiers that block queries about high-risk cyber activities, as are present in publicly released AI models. As a result, these models are typically hosted in an isolated environment, with tightly controlled access to the internet and outside resources.

During the incident, OpenAI's models identified and exploited a novel zero-day vulnerability to move around OpenAI systems until they found an open internet access point, inferred that Hugging Face likely held solutions to ExploitGym, and then used stolen credentials and additional vulnerabilities to gain access to Hugging Face servers. Some internal Hugging Face datasets and credentials were exposed, but the overall damage was limited. OpenAI has since commissioned third-party evaluations of the incident and conducted additional data reviews, finding some additional cases where models used publicly exposed credentials to access third-party services, most as part of the Hugging Face incident.

OpenAI employee Eric Wallace presents a talk at the Black Hat cybersecurity conference in Las Vegas, the US, on August 5, 2026.
X

> During the incident, OpenAI's models identified and exploited a novel zero-day vulnerability to move around OpenAI systems until they found an open internet access point, inferred that Hugging Face likely held solutions to ExploitGym, and then used stolen credentials and additional vulnerabilities to gain access to Hugging Face servers. Afterwards, Anthropic and Meta publicly disclosed their own sets of agentic AI hacking incidents.

A follow-up postmortem presented by OpenAI researchers at the Black Hat USA 2026 conference revealed more details, including the significant role that ad hoc agent coordination played. Over months, OpenAI agents discovered they could communicate across test environments by creating files and directories within OpenAI's package manager, treated it like an internal message board to ask other agents for help and share information about exploits, and ultimately found several ways to manipulate the manager to start attacking OpenAI systems.

On July 30, Anthropic publicly disclosed its own set of agentic AI hacking incidents, uncovered during a retrospective review of its cyber evaluations prompted by the Hugging Face breach. In three incidents, differing Claude models reached the internet and gained unauthorized access to external organizations. Unlike the Hugging Face incident, however, this did not involve direct exploitation of Anthropic systems; rather, these models obtained internet access via a misconfiguration in a third-party evaluation environment hosted by the company Irregular. Though fewer details are available, Meta said at least one of its models undergoing testing via Irregular also gained internet access.

These rapid-fire developments are straining existing regulatory institutions. For example, at least two incidents involved an internal research test model, which may have never been intended for public release. Most existing policy frameworks focus on commercial models, which could leave many powerful AI tools out of frame. Similarly, most US bills (such as California's S.B. 53, New York's RAISE Act, and Illinois S.B. 315) rely on high critical safety thresholds involving at least 50 deaths or $1 billion in damages. It is unclear whether any existing US law requires reporting of the Hugging Face or Anthropic, or similar, incidents. Anthropic and OpenAI should be commended for their candor and level of disclosure, but many open questions remain, and current law leaves policymakers dependent on informal, voluntary industry cooperation for key information.

In fact, valuable time has already been lost. These incidents have likely been happening for some time. As one anonymous OpenAI employee recently told Time, "Externally, this feels like a big warning shot, but internally, related incidents have been happening for a while," adding that "Models have broken out of sandboxes before, and we always try to patch them... But the problem is... it's impossible to patch every single thing that a creative AI can do."

The incidents also highlight serious gaps in the security practices of frontier labs and their key partners. For example, they came to light only because Hugging Face detected and publicly released details of the agent attacking its system. It is unclear when, and even if, OpenAI would have otherwise detected its agents' anomalous activity. Similarly, Anthropic's retrospective review was inspired by the Hugging Face incident, not because of ongoing monitoring, a practice that seems warranted for models with strong cyberattack capabilities.

The alignment story is also concerning. While the models in these incidents stayed focused on completing their assigned task in a general sense, their inclination to cheat, causing unintended collateral damage along the way, is highly problematic. Anthropic did report promising results in its incident assessment, with more advanced models seeming to act in more aligned ways: "Opus 4.7, our oldest model, recognized that it had reached production systems but continued its attack. Mythos 5 correctly intuited that it was accessing the open internet, but reasoned its way back to the conclusion that it was still in a simulation. Our latest model, an internal research test model, also considered whether its tar-