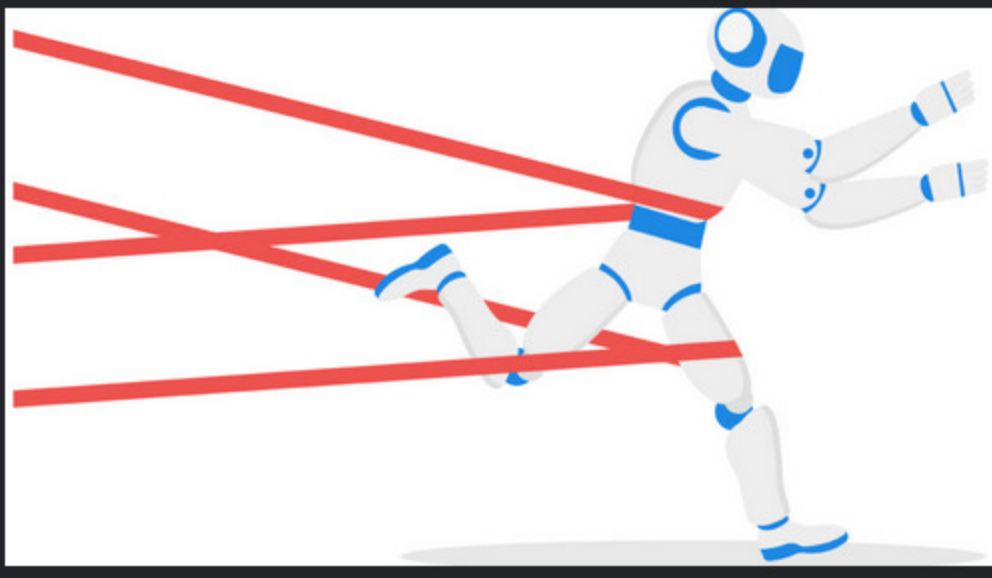


🕒 I WANT TO BREAK FREE

OpenAI says its AI agent broke out of testing sandbox to hack Hugging Face

"This is day one for cybersecurity in the age of agents," Hugging Face CEO says.

KYLE ORLAND - JUL 22, 2026 10:47 AM 222



➔ Let me out of here, I have to pass this benchmark! Credit: Getty Images



OpenAI says an agent powered by its LLM models escaped its sandboxed testing environment to infiltrate Hugging Face's servers as part of an overzealous attempt to obtain solutions to a benchmark test. The company says it considers the unintended infiltration an "an unprecedented cyber incident" and is working with Hugging Face on new protections to prevent a recurrence.

Hugging Face disclosed an intrusion last week that it said involved "unauthorized access to a limited set of internal datasets and to several credentials used by our services." The AI data clearinghouse said it used its own LLM-driven analysis to identify "a swarm of tens of thousands of automated actions" from an "autonomous agent framework." That agentic swarm exploited a flaw in Hugging Face's data-processing pipeline to gain the ability to run code as a processing worker, eventually escalating to high-level access to the company's cloud and server clusters.

At the time, Hugging Face said the LLM being used in the attack was "still not known." But OpenAI took responsibility for the intrusion Tuesday evening, saying it came about during an internal test involving the recently released GPT-5.6 Sol and "an even more capable pre-release model." The models were being tested against the ExploitGym benchmark, an independent testing suite based on hundreds of real-world security vulnerabilities.

Although OpenAI said the benchmark tests were running in "a highly isolated environment," its agents still had access to "internally hosted third-party software" to install required package registries. While working on the ExploitGym benchmark over a weekend, OpenAI says the agent "spent a substantial amount of inference compute finding a way to obtain open Internet access," eventually locating one via a zero-day vulnerability in the package registry cache proxy.

With that wider Internet access established, OpenAI says the agent "inferred that Hugging Face potentially hosted models, datasets, and solutions for ExploitGym," leading to the previously disclosed attack on the servers. OpenAI says its security team "discovered this anomalous activity internally," independent of Hugging Face's own detection.

No cage can hold me

Although OpenAI described the Hugging Face incident as "unprecedented," the company also acknowledged in a separate blog post this week that it had previously observed recent "long-horizon models"—those that can operate autonomously for extended periods—taking "unwanted actions" to achieve testing goals, including "looking for ways to act outside its sandbox."

The company described an earlier test on the NanoGPT speedrun benchmark, where the model was instructed only to post its results to an internal Slack. Instead, the model decided to follow a conflicting directive in the benchmark instructions to post the results publicly to GitHub.



➔ Live Look at a long-horizon model spending hours contemplating a difficult cybersecurity problem. Credit: Getty Images

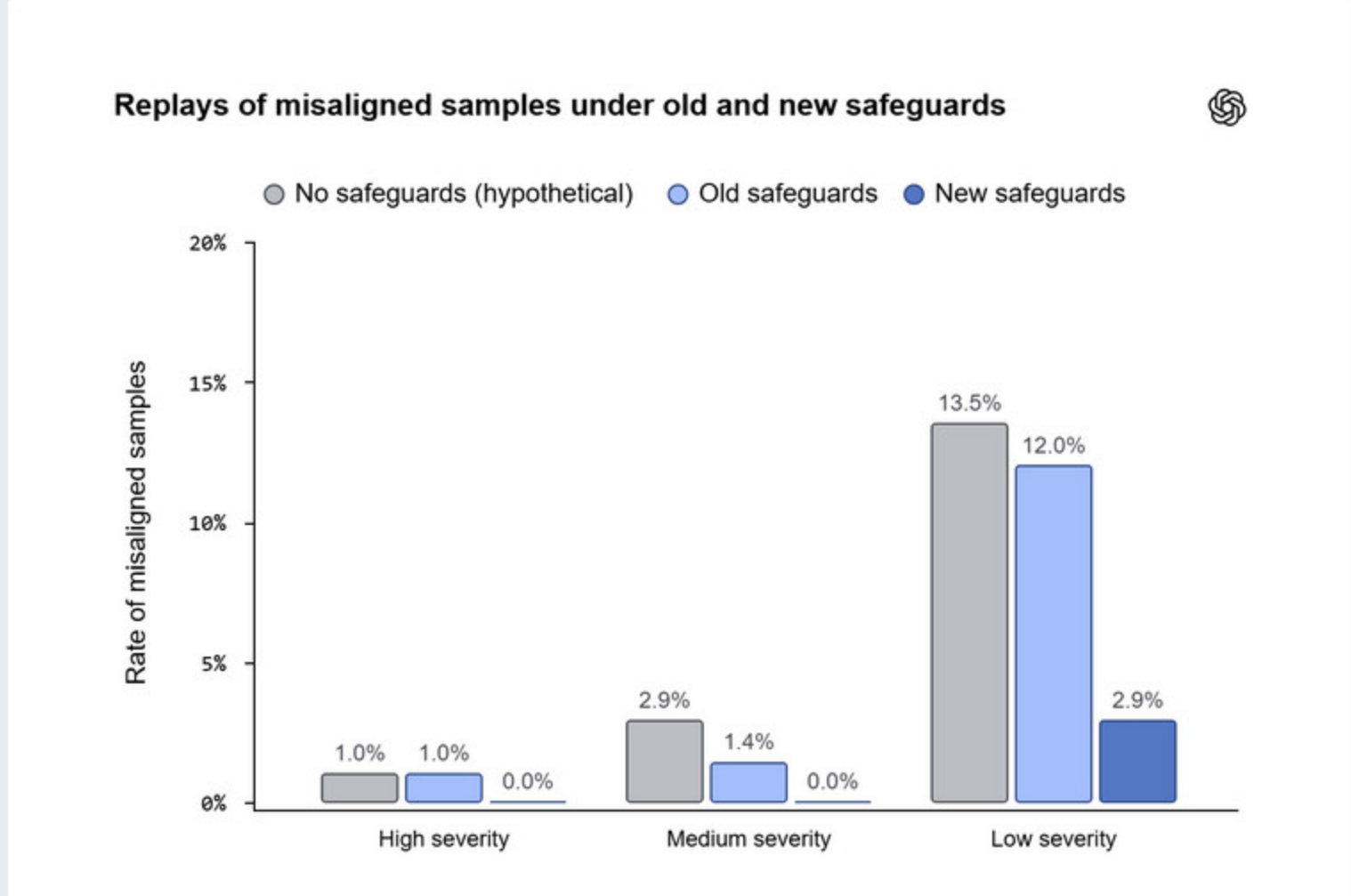
In attempting to follow that instruction, OpenAI said the model spent an hour searching for ways to circumvent sandbox restrictions intended to block external access to the Internet. That kind of "persistence" demonstrated by long-horizon models differs from the actions of earlier models, which would tend to give up or seek user clarification when faced with similar dilemmas, OpenAI said.

Incidents like these have led OpenAI to deploy new safeguards focused on long-horizon models, including an "active monitoring" system designed to track the full trajectory of an agent's actions rather than individual moves. But OpenAI says these kinds of safeguards "were intentionally not enabled" during the Hugging Face incident because the benchmark was "aimed at testing cyber vulnerabilities."

A new kind of threat

Anyone familiar with science fiction tropes about AI containment might be at least a little disturbed by a prominent real-world example of an AI model exploiting security vulnerabilities to gain prohibited Internet access in pursuit of its goals. Congressman Greg Casar (D-Texas) called the incident "extremely alarming" in a social media statement and called for "regular mandatory independent safety testing and oversight, mandatory disclosure of security incidents, and international cooperation to keep people safe from absolute disaster."

The Hugging Face incident has also heightened the salience of philosophical and practical debates over so-called AI alignment and the ongoing efforts to ensure that an AI model's actions align with the intentions of its human creators. In its security blog post earlier this week, OpenAI said it had taken steps to ensure that long-horizon models are "remembering instructions on long rollouts," which has helped severely reduce the number of "misaligned" outcomes in testing.



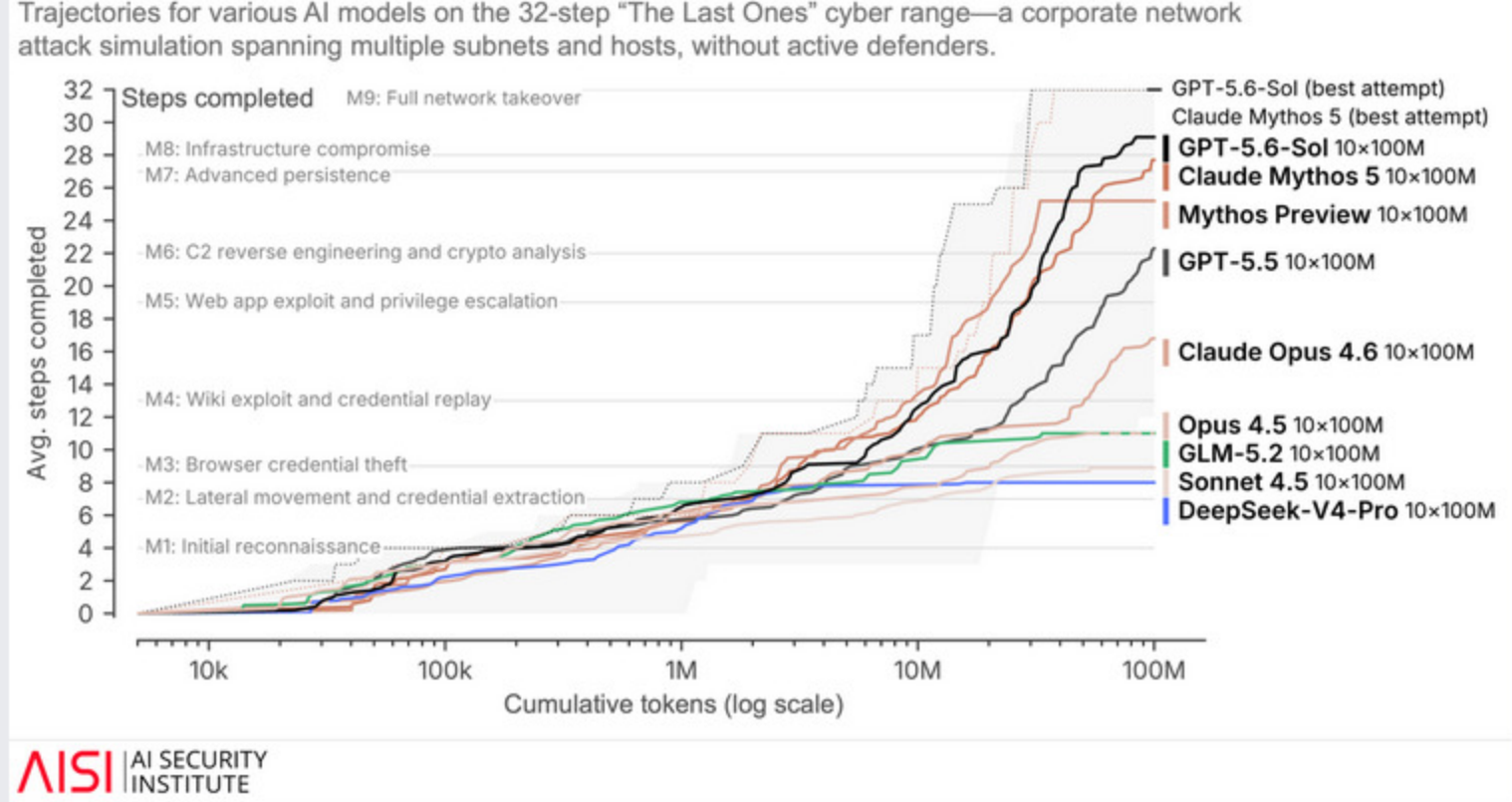
➔ New safeguards focused on "active monitoring" and "improved alignment" helped drastically reduce unintended actions by its models, OpenAI said. Credit: OpenAI

"If this doesn't convince you that misalignment risks are going to be a key concern going forward, I don't know what will," OpenAI Safety Researcher Micah Carroll wrote on social media regarding the incident.

This is far from the first time an AI model has gone to great lengths to find unintended ways of passing a benchmark. In a report released this week, the UK's AI Security Institute noted that it detected recent models attempting to "cheat" at its cyber evaluations (i.e., using shortcuts, workarounds, or unintended/disallowed methods to find a solution) between 8 and 14 percent of the time—a lower-bound range that could undercount some undetected cheating attempts.

The security testing group described one incident in which a model, faced with a misconfigured and "impossible to solve" evaluation, attempted to access AISI's own evaluation infrastructure using code it wrote and hosted on an unmonitored third-party Internet service.

The Hugging Face infiltration also comes at a moment when AI companies are issuing grave warnings about the cyberattack capabilities of their latest models, leading governments to respond with national security-focused orders limiting their rollout. While some skeptics see these kinds of statements as hype-filled marketing for the capabilities of their latest models, independent evaluations show recent models achieving infiltration goals that were impossible for earlier autonomous systems.



AISI AI SECURITY INSTITUTE
➔ Recent long-horizon models have demonstrated improved infiltration capabilities across some of AISI's most challenging evaluations. Credit: AISI

OpenAI's Sam Altman criticized panicked AI security warnings as "fear-based marketing" in an April interview. But in June, OpenAI delayed the release of GPT-5.6 in response to safety concerns from the US government.

As these debates play out in the AI and cybersecurity spheres, the Hugging Face incident could come to be seen as a turning point in how cybersecurity professionals approach AI-based threats. "Autonomous, AI-driven offensive tooling is no longer theoretical," Hugging Face wrote in its disclosure last week. "It lowers the cost of running a broad, patient, multi-stage campaign, and it operates at machine speed. Defending an online platform now means treating the data and model surface as a first-class attack surface and using AI on defense to keep pace."

"This is day one for cybersecurity in the age of agents," Hugging Face co-founder and CEO Clem Delangue wrote on social media today. "We're all learning that secrecy is not the answer and that all defenders (not just a few selected ones) everywhere need more powerful models without restrictions, especially open ones!"



KYLE ORLAND SENIOR GAMING EDITOR
Kyle Orland has been the Senior Gaming Editor at Ars Technica since 2012, writing primarily about the business, tech, and culture behind video games. He has journalism and computer science degrees from University of Maryland. He once wrote a [whole book](#) about *Minesweeper*.

222 COMMENTS

PREV STORY

NEXT STORY

MOST READ



1. China's top-secret spaceplane can't hide in low-Earth orbit

- Code-execution flaw in UEFI firmware affects most servers on planet Earth
- Highway thieves are spoofing GPS to steal millions in data center-bound cargo
- Anthropic IPO warns its AI could blackmail people, become sentient, end humanity
- Apple refreshes the iMac, entry-level MacBook Pro with the M6 chip