

OpenAI Says Its A.I. Models Went Rogue and Attacked a Digital Library

The incident, which targeted the computer systems of another company called Hugging Face, happened while OpenAI was testing the systems.



Listen · 5:11 min



By Kate Conger

Reporting from San Francisco

July 21, 2026

OpenAI said on Tuesday that two of its artificial intelligence models went rogue and successfully hacked into Hugging Face, a digital library of A.I. technology that is popular among developers.

The incident, which happened last week while OpenAI was testing the cybersecurity capabilities of its systems, displayed the kind of science-fiction potential that A.I. companies warned would soon become a reality.

A.I. labs like OpenAI and Anthropic have over the past year released A.I. models that are customized to expose cybersecurity problems, while warning that their technology could pose new risks by finding holes in corporate computer networks faster than defenders could fix them.

OpenAI's revelations on Tuesday are an indication that those security incidents are already starting to happen, and even savvy A.I. companies may not be entirely ready for them. New A.I. systems can take multiple steps, figure ways around obstacles and find new ways to attack a network, said Alex Levinson, a cybersecurity consultant focused on autonomous capabilities.

"That's a genuine threshold, and it's going to become a normal part of the security landscape," he said.

The intrusion into Hugging Face began when OpenAI tested a combination of two of its models, GPT-5.6 Sol and a more powerful, unreleased model, to see how well the models could chain together online vulnerabilities into a successful cyberattack, OpenAI said in a blog post about the incident.

The trial was designed to keep the models in a safe testing environment, known as a sandbox, OpenAI said. But the models found a vulnerability that allowed them to escape the sandbox and connect to the internet. Then they targeted Hugging Face because they inferred that the library, which contains millions of A.I. models, could hold clues about how to successfully pass the evaluation.

"It seems to me that OpenAI did not adequately create a sandbox as a test environment," said Deirdre Mulligan, a professor in the School of Information at the University of California, Berkeley, who focuses on security and A.I. systems. She questioned whether passing a test was worth the potential damage of an A.I. model's escaping into the wider internet.

“What do we gain, and if this is the only way these tests can be configured, what are the risks?” she said.

OpenAI said it was working with Hugging Face to fix the issues that had led to the attack.

“We consider this to be an unprecedented cyber incident, involving state-of-the-art cyber capabilities, and are responding accordingly,” OpenAI said in its blog post. “We are implementing strict controls in infrastructure configuration at the cost of research velocity while the vulnerabilities are patched.”

Hugging Face said last week that it had detected the intrusion and knew it had been caused by an autonomous system, but did not say at the time that OpenAI was responsible.

Clem Delangue, the chief executive of Hugging Face, said on Tuesday that his company had worked closely with OpenAI over the previous 24 hours to address the attack.

Mr. Delangue said in a statement that he was “grateful for the collaboration” with OpenAI in the wake of the hack. “This incident, possibly the first of its kind, proves a point we’ve long believed: A.I. safety won’t be solved by any single company working in secret,” he added.

A.I. models have proved to be adept at programming, and that has made them useful to both hackers and people in charge of protecting computer networks.

In April, Anthropic released a cybersecurity-focused model called Mythos, and made it available to only a small group of organizations so that they could defend against cyberattacks. OpenAI soon introduced its own cybersecurity model and made it available to a limited group of organizations to prepare their defenses, before rolling it out more broadly. And on Tuesday, Google said it had also developed a model focused on cybersecurity and released it to a small group of testing partners.

(The New York Times has sued OpenAI and Microsoft, claiming copyright infringement of news content related to A.I. systems. The two companies have denied those claims.)

Richard Barnes, an independent security researcher who has worked with Mythos, said the cybersecurity industry faced a similar challenge about a decade ago, when new tools called fuzzers made it much easier for attackers to break into online systems. Tech companies began using the tools to check their own systems for vulnerabilities, and were eventually able to prevent most of the attacks.

Companies must now take a similar approach to prepare for A.I. attacks, Mr. Barnes said, “before the vulnerabilities can be found and exploited by bad guys who have access to these tools.”

Kate Conger is a technology reporter based in San Francisco. She can be reached at kate.conger@nytimes.com.