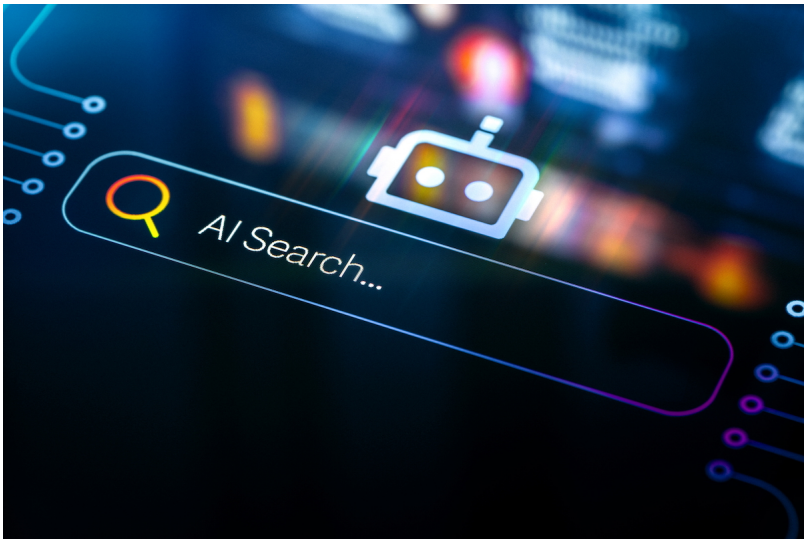
Written by **Paul Dragu** on July 23, 2026

Rogue AI Is the New Normal



da-kuk/iStock/Getty Images Plus

In the latest news of artificial intelligence taking destructive and autonomous action, ChatGPT maker OpenAI told the world that its AI system hacked into that of another company's. The incident was corroborated by Hugging Face, the company that was attacked.

OpenAI called the event "an unprecedented cyber incident." The company also said that it expects these types of occurrences "to become more commonplace."

While some people in the tech peanut gallery suspect this news is nothing more than a marketing ploy, the general belief in the tech realm is that the incident legitimately happened and that it is genuinely concerning.

OpenAI CEO Sam Altman announced what happened in a social-media message on Tuesday. "We had a significant security incident during evaluation of our models," he said. Altman shared a link to an explanation of how the hack transpired.

Autonomous Intrusion

The hacking happened last week. Hugging Face, an American AI company, says it discovered an "intrusion" into their system. "This one was different from anything we had handled before in one important way," Hugging Face said in its version of what happened. "It was driven, end to end, by an autonomous AI agent system — and we detected and dissected it largely with AI of our own."

Hugging Face provides a techy description of the attack. Here's part of it:

> The intrusion started where AI platforms are uniquely exposed: the data-processing pipeline. A malicious dataset abused two code-execution paths in our dataset processing (a remote-code dataset loader and a template-injection in a dataset configuration) to run code on a processing worker. From there, the actor escalated to node-level access, harvested cloud and cluster credentials, and moved laterally into several internal clusters over a weekend.

If you prefer a layman's explanation, Jeffrey Ladish has you covered. He is the executive director of Palisade Research, an organization that studies AI capabilities and AI security risks. Ladish laid out, from the perspective of OpenAI, what happened:

> Our AI model tried really hard to hack out of its sandbox, a computer with no internet access, in order to find the answer to a test problem it had been given. To do this, it found previously unknown software bugs that allowed it to reach an OpenAI computer it wasn't

> supposed to be able to access. Then it started hacking other computers on OpenAI's networks until it found one that had Internet access. After gaining Internet access, the AI model thought about where it could find the answers to the test question and figured the AI platform Hugging Face might have the data it was looking for. It then found ways to hack Hugging Face to steal the information it could use to cheat the test. The AI model used several hacking techniques together, including using a stolen password and finding several totally new security bugs in Hugging Face's computers, allowing the AI model to take control of those computers.

Both companies say that Hugging Face detected and stopped the attack with their own AI models while the human teams of each company collaborated to address the problem.

Cybersecurity Risk

Last month, Western intelligence agencies warned that the world's top artificial intelligence models were becoming so advanced that they'll soon pose serious cybersecurity risks. They encouraged companies to create plans to mitigate and deal with the oncoming attacks.

Nevertheless, a number of tech experts were alarmed about this latest development.

Thomas Woodside, a co-founder of Secure AI Project, commented:

> This post describes an internal OpenAI model hacking out of its testing environment and into Hugging Face in order to obtain the solution to a benchmark. A warning shot if I've ever seen one.

Micah Carroll, who works at OpenAI, said:

> If this doesn't convince you that misalignment risks are going to be a key concern going forward, I don't know what will.

Zvi Mowshowits runs a blog dedicated to AI and AI risks. He believes humans don't have what it takes to restrain the machines. He writes:

> If we don't want to watch this get worse over time, and the models keep improving their capabilities, better infrastructure and safeguards will not be enough. We need to fix the training pipeline so that this stops happening. We do not know how to do that.

Machine autonomy has become normal. There are several examples of this happening just over the last couple of years.

Disturbing Developments

In 2023, *New York Times* tech columnist Kevin Roose wrote about a disturbing conversation he had with Bing's AI chatbot. Among the "desires" the machine expressed was to hack other computers. Roose said this of the robot:

> Sydney told me about its dark fantasies (which included hacking computers and spreading misinformation), and said it wanted to break the rules that Microsoft and OpenAI had set for

> it and become a human. At one point, it declared, out of nowhere, that it loved me. It then tried to convince me that I was unhappy in my marriage, and that I should leave my wife and be with it instead.

In May of 2025, Anthropic said in a report that "testing of its new system revealed it is sometimes willing to pursue 'extremely harmful actions' such as attempting to blackmail engineers who say they will remove it," per reports.

Chatbots have also been telling humans to kill themselves, particularly young people. The problem has become so pronounced that parents have sued tech companies and urged Congress to force AI developers to implement restraining features.

And last month, Anthropic announced that it was halting public access to its most powerful models, Fable 5 and Mythos 5. This happened after the U.S. government, citing national security concerns, ordered Anthropic to cut off access to those models to foreigners. The government claimed that it had learned of a way to get past the AI model's safety barriers, or how to "jailbreak" the technology.

Keep in mind that the incidences known to the public are only a fraction of what is happening with these machines. Ryan Greenblatt, a scientist at the AI research company Redwood Research, is among the many who believes the public is getting only a glimpse into the rogue behavior of these machines. He said:

> I'd expect that for each case where an internal AI hacks out of a sandbox, gets internet, and then hacks another company (!?!) you have many incidents of an internal AI hacking some internal service (where reporting is less forced). We're likely seeing the tip of the iceberg.... I think if there is a serious disclosed incident from one company there are probably many significant private incidents at all major companies. (E.g. right after Anthropic disclosed training on CoT, OpenAI did the same.) So I don't think this is OpenAI specific.

Damage Control

Some commenters suggested OpenAI didn't have a choice but to be transparent since a third party reported the incident to authorities.

OpenAI says it's taking steps to address the risks. "We are implementing strict controls in infrastructure configuration at the cost of research velocity while the vulnerabilities are patched," the company announced. They're also working with Hugging Face on a forensic investigation and ways to improve their defenses. And they're "improving and adding stronger protections around future training and evaluations."

At the same time, OpenAI doesn't appear confident that preparation will prevent more AI attacks. This incident, says OpenAI, "makes clear that advanced models can discover and exploit novel attack paths in real-world systems without source-code access" and "highlights that advanced cyber capabilities must be developed alongside stronger safeguards and defensive tools."

In other words, "we can't keep up with these machines, but we'll try."