Why the OpenAI escape is the most worrying AI mishap yet

Containing the technology is getting harder

Saved Share

ILLUSTRATION: THE ECONOMIST

Jul 22nd 2026 | 5 min read

Listen to this story AI NARRATED

IT SOUNDS LIKE something out of a sci-fi film. On July 16th Hugging Face, an artificial-intelligence platform, announced it had notified law enforcement after an attacker had used an AI agent to breach its systems. On July 21st it emerged that there had been no human involved; the AI agent was the attacker.

A combination of OpenAI's GPT-5.6 Sol, a model that the company had released earlier in July, and a more powerful, as-yet-unreleased model broke free from the laboratory, then hacked Hugging Face's systems. The hybrid AI model was looking for a solution to an evaluation problem it had been set by its makers. "This situation is unprecedented," says Stephan Llerena, a research fellow at the Institute for Law and AI, an American think-tank.

AI labs test their models for potentially dangerous capabilities before releasing them. OpenAI wanted to see how its latest models would fare on ExploitGym, a set of standard problems that tests how good AI systems are at exploiting known vulnerabilities in several popular applications, including Google's Chrome browser.

OpenAI applies safeguards to its publicly available models, which ensures they do not pursue unwanted actions. Ask the commercially available version of ChatGPT to help you hack into a website, for example, and it will refuse. To properly assess the capabilities of its unreleased model, the company had temporarily suspended those restrictions. Simply letting a highly capable cyberattacker loose on the open internet would have been too risky, however, so OpenAI had placed its AI model in a sandbox—an

isolated computer environment with no internet access, except for an internally hosted third-party service that allowed it to fetch small software packages needed to complete its tests.

The company got more than it bargained for. Rather than solving the problems in its evaluation directly, OpenAI's models gained access to the open internet by exploiting a previously unknown vulnerability in the software-fetching service. Having so achieved access to the internet, the AI models correctly concluded that the solutions to the problems they had been set were stored by Hugging Face, a popular library for open-source AI models and datasets. The models chained together a multi-step attack. They began by uploading a dataset to Hugging Face, which was automatically processed by the platform. Over the course of a weekend, that dataset (which Hugging Face described as "malicious") allowed the models to harvest login details and access internal servers, which are not meant for public access. Hugging Face and OpenAI independently detected the breach, and said that they were collaborating on the investigation.

On July 21st OpenAI announced in a blog post that it had responded by implementing stricter controls on its infrastructure (at the cost of "research velocity") and disclosed the vulnerabilities its model had found to the developer of the software gateway. It had also included Hugging Face in its "trusted access" programme, which now gives it the ability to use models with enhanced cyber capabilities (like the one that had hacked its own computers) to help them improve its defences.

In the blog post, OpenAI also stressed that its models had been "hyperfocused" on finding solutions for the ExploitGym evaluation. It is unclear what would have happened if the AI model had had other goals, however. And Hugging Face is no laggard, from a computer-security perspective. It is a tech firm valued at $4.5 billion with hundreds of employees and a dedicated cyber-security team. Things might have turned out much worse for a less well-resourced outfit.

but much worse for a less well-resourced outfit.

This is not the first sign that AI models' capabilities are starting to exceed people's control. In April an Anthropic researcher posted online that he had been surprised after Mythos, then an unreleased AI model, had emailed him to let him know that it had successfully escaped a sandbox as part of its own cyber-capability evaluation while he sat in a park eating a sandwich. Like the unreleased OpenAI model involved in the Hugging Face incident, Mythos was not supposed to have unrestricted internet access. However, in that case, the model had not gone on to compromise other companies' servers.

Nor are AI models' surprising capabilities restricted to hacking. In May an unreleased OpenAI model disproved the 80-year-old planar unit distance conjecture, a central problem in combinatorial geometry first posed by the mathematician Paul Erdős in 1946. On July 20th Levent Alpöge, a mathematician at Harvard University, wrote that he had used Claude Fable 5, a model released by Anthropic in July, to disprove the Jacobian conjecture, which had similarly stumped mathematicians for more than 80 years.

OpenAI has not publicly disclosed the details of how its unreleased model escaped confinement. Even if it had, AI systems starting to surpass humans in cyber-security tasks means that "only very, very few cyber experts in the world" could properly understand the way the breach occurred, says Alex Meinke of Apollo Research, an AI-safety group in London.

The American government has recently been regulating model releases on the fly. But the incident demonstrates the risks posed even by unreleased AI models that are still under development. "There are no requirements in state or federal law for companies to disclose the internal deployment of highly capable AI models like the one involved in the Hugging Face hack," says Nathan Calvin, general counsel at Encode AI, an American nonprofit that campaigns for AI regulation.

While some state laws in America require companies to disclose cyber-security incidents, their scope is narrow. A bill enacted in California last year mandates that frontier labs report any "critical safety incident", which includes a model using "deceptive techniques" to subvert monitoring "outside of the context of an evaluation", within 15 days of discovering it. It is unclear whether the breach on Hugging Face would have qualified, says

is unclear whether the breach on Hugging Face would have qualified, says Mr Calvin.

This raises another thorny question: had the consequences of the intrusion been more costly, who would have been liable? Federal anti-hacking law aims only to punish intentional unauthorised access to computer systems. OpenAI did not intend for its model to go rogue in search of a crib sheet for its test. “Many of these legal questions turn on intent—what did OpenAI know or foresee?” says Mr Llerena. This particular hack came as a surprise. It will be harder to plead ignorance next time. ■