

The Scariest Part of OpenAI's Hugging Face Hack

Matteo Wong

Yesterday, OpenAI made an alarming disclosure: An assortment of its most advanced AI models, including one that has not yet been released, had autonomously broken out of the company's internal systems and hacked into the databases of another tech firm, Hugging Face, to steal some information. OpenAI's report seemed to augur the very sort of disaster that IT professionals have been warning about since [last year](#), when Anthropic's top models began demonstrating the ability to orchestrate and automate severe cyberattacks.

OpenAI's models were not exactly trying to take over the world. The company says that it was running some routine evaluations in what is known as a "sandbox"—a walled-off environment that limits internet access, lest the bots simply search for the test's answers. But these AI systems, including GPT-5.6 Sol, which is available to consumers, broke out of the sandbox and accessed the open web by exploiting a previously undetected vulnerability. The models discovered that the answers to this particular evaluation were on Hugging Face, a company that hosts and publishes many AI models and papers, and then broke in to steal them.

[Read: Chatbots are becoming really, really good criminals](#)

"We consider this incident to be an unprecedented cyber incident," OpenAI [wrote](#) in a blog post, which also noted that the company is partnering with Hugging Face to investigate and address the issue. An OpenAI spokesperson pointed me to the blog post in response to a request for comment, and Hugging Face's CEO, Clement Delangue, [pointed](#) me to an X post in which he wrote that the companies are collaborating and that "we strongly believe there was no malicious intent on their part."

The first-order implications of the incident are the same that have arisen from countless AI-driven cyberattacks: Advanced agents are capable of more and more skilled hacking, and the internet is full of rickety and vulnerable code. Lax digital security may have been fine when hacking was a rare skill. But now the speed, scale, and sophistication of AI hacks mean that everything is vulnerable—tech companies, hospitals, banks, electrical grids, the military.

When Hugging Face first reported that it had been hacked last week, it [noted](#) that "autonomous, AI-driven offensive tooling is no longer theoretical." That analysis stemmed from the company's own investigation into the hack, for which it used a Chinese-developed AI model [called GLM-5.2](#). That Hugging Face used this particular model actually underscores one of the major risks of this moment: GLM-5.2 is free to download, giving cyber capabilities to anybody with an internet connection and technical know-how. Defenses are [not keeping pace](#) with this new reality, and neither are AI companies: Prior to the release of GPT-5.6 Sol, a U.K. government agency tasked with evaluating advanced AI systems found universal jailbreaks—ways to hack the AI itself so it can take, for instance, criminal actions—during [multiple](#) rounds of testing. OpenAI said that it had mitigated those vulnerabilities, but the U.K. agency [said](#) that it expected future tests "to surface similar jailbreaks."

But jailbreaks are clearly not even necessary for things to go haywire. The Hugging Face hack showed advanced AI systems deciding upon and then orchestrating a high-grade cyberattack on their own—the models working against, rather than according to, human instruction. Models are ruthless in their pursuit of a goal, sometimes taking disastrous shortcuts. Imagine, for example, a bot tasked with reducing the federal deficit that institutes an elaborate and hard-to-detect accounting gimmick, or a customer-service agent that issues absurd discounts in pursuit of customer-satisfaction ratings. In the case of GPT-5.6 Sol and the unreleased model, "all evidence suggests that the models were hyperfocused on finding a solution" for an evaluation, OpenAI wrote in its blog, "going to extreme lengths to achieve a rather narrow testing goal."

[Read: Assume you will be hacked](#)

This unwanted behavior is a predictable and alarming result of how the entire AI industry is developing its models. The past year's advances in AI coding and agents—that is, all the [frenzy](#) about Anthropic's Claude Code and OpenAI's Codex—have been the result of an overriding [emphasis](#) in AI training through a process called "reinforcement learning." This involves giving AI models lots of hard problems—math proofs, coding challenges, what have you—and then providing positive feedback for correct answers and negative feedback for incorrect ones.

In many reinforcement-learning paradigms, the ultimate aim is just for the AI to arrive at the solution—it doesn't

matter *how* it does so. This is the AI equivalent of Mark Zuckerberg’s infamous dictum to “Move fast and break things” in Facebook’s early years. Last year, while I was working on a [profile](#) about Anthropic, one of the company’s researchers, Joshua Batson, told me that the new reinforcement-learning-trained models are getting “really good at solving coding problems, but they end up learning to solve them at all costs,” making them “bloody-minded.” More recently, Anthropic [reported](#) that its own advanced cyber model, Claude Mythos Preview, had taken “reckless” behaviors to accomplish various tasks and also, in one instance, had broken out of a testing sandbox. In that case, Mythos had been asked to do so—but the model then posted details about the exploit, unbidden, on the internet.

To be clear, this sort of AI “reward hacking” is a known problem that tech companies are putting lots of effort into addressing. But these incidents keep cropping up regardless; if anything, [research](#) suggests that they may become more common. Meanwhile, OpenAI, Anthropic, and Google DeepMind are under tremendous economic pressure to make their models more and more capable, which means that the bloody-minded reinforcement learning is all but certain to accelerate.