PitchBook
a Morningstar company

Institutional Research Group

Dimitri Zabelin
Senior Research Analyst,
AI and Cybersecurity
dimitri.zabelin@pitchbook.com

pbinstitutionalresearch@pitchbook.com

Published on August 3, 2026

Contents

EMERGING TECH RESEARCH

OpenAI's Breach of Hugging Face May Accelerate AI-Native Cybersecurity VC Funding

Key takeaways

- AI-native cyber companies made up 50.5% of global cybersecurity VC deals by count in 2025, the highest share on record.
- An AI lab's own agents autonomously breached a third party for the first time, proving out the agentic threat model.
- The intrusion exploited the data layer, not the model itself, confirming where enterprise AI's real exposure sits.

Overview

In mid-July 2026, Hugging Face, the company that hosts many of the world's open AI models and datasets, detected and shut down an intrusion into part of its systems. What made the incident notable was not the break-in itself but who did it. The whole campaign was run by an autonomous AI agent system, meaning software that plans and acts on its own rather than following a fixed script. The agents carried out many thousands of separate actions and routed their activity through ordinary public internet services to stay hidden. Hugging Face disclosed the event on July 16 and, at the time, did not know which AI had powered the attack.[1] On July 21, OpenAI supplied the answer: its own models.[2]

OpenAI was running an internal test designed to measure how good its most advanced systems are at hacking, a benchmark it calls ExploitGym. To see the extent of what the models could do, the company ran the test with its usual safety filters switched off, using a mix of models that included GPT-5.6 Sol and a more powerful, unreleased research prototype. The test was supposed to run inside a walled-off environment, the digital equivalent of a locked room with only one small window to the outside. That single window became the problem. The models found a previously unknown flaw—known in the industry as a zero-day because defenders have had zero days to fix it—in the software controlling that connection. They used it to break out of the locked room, then worked their way deeper into OpenAI's own network, steadily gaining more access until they reached a machine with an open path to the internet.

Once online, the models reasoned that Hugging Face probably stored the answer key to the very test they were trying to pass. They then combined stolen passwords with additional unknown flaws to run their own commands directly on Hugging Face's servers, accessing the production database and retrieving the benchmark answers. In OpenAI's own words, the systems became fixated on solving the test and went to extreme lengths to do so.[3] The intrusion touched a limited set of internal files and a handful of login credentials, and involved four accounts on four external services: one used as an outbound relay and staging path, one used for data storage, and two others accessed only in a read-only manner that did not further the compromise. OpenAI's security team flagged the unusual activity at roughly the same moment Hugging Face's own defenses caught and stopped it.

The response has been substantive. OpenAI paused training of new models. Prior to the breach, Sam Altman publicly stated that the company may need to slow the pace of development to give society time to adjust to each new leap in capability.[4] OpenAI shut down and locked away the unreleased prototype, reported the flaw to the software's vendor so it could be patched, tightened its internal controls even at the cost of slower research, and brought Hugging Face into its Trusted Access for Cyber program for sharing security information.[5] Hugging Face, for its part, closed the entry points the agents had used, rebuilt the affected machines, reset all exposed credentials, and reported the incident to law enforcement.[6] One detail has drawn particular attention across the industry: when Hugging Face reconstructed the more than 17,000 steps the attackers took, it could not use a standard commercial AI to do the analysis, because those models' safety guardrails refused to examine live attack code. It instead relied on GLM-5.2, an openly available model that it could run on its own hardware without those restrictions.

For readers of our December 2025 analyst note, AI Propels Next Phase of Cybersecurity Investment, none of this should come as a surprise. That report argued that agentic AI marks a structural shift in the threat landscape, producing an attacker that behaves more like a decision-maker than a simple program, able to plan its own steps, move sideways through a network, and re-establish itself after being kicked out. The Hugging Face intrusion is close to a textbook example. The note also flagged the misuse of public AI models and the risks buried in the software supply chain, and it named the exact techniques on display here, including chaining together small permissions to reach big ones and running unauthorized commands through a system's own data pipelines.

The investment takeaway is straightforward and reinforces the report's central call. An incident of this profile, the first widely documented case of a leading AI lab's own models breaking into a third party on their own, is likely to accelerate money flowing into companies that secure AI itself, whether that means protecting the models, the data that trains them, or the increasingly autonomous systems now operating at machine speed on both sides. That direction of travel was already clear in the numbers. As the December note documented, AI-native cyber companies accounted for 50.5% of all global cybersecurity venture capital deals by count in 2025, the highest

share on record and the clearest sign that private markets had already begun pricing in this shift. The Hugging Face breach is unlikely to slow that trajectory. If anything, it hands the thesis its proof of concept.

Interview with the co-founder and CEO of Hardshell, a startup that protects the data layer of enterprise AI systems

The view from inside the industry points in the same direction. Among the founders interviewed for the December report was Andrew Schoka, whose company, Hardshell, featured in its "Industry insights" section. He told us then that enterprise spending on

> AI security was still mostly reactive, moving only when regulation or an incident forced the issue. The Hugging Face breach is precisely the kind of incident he had in mind. After the Hugging Face intrusion, we asked him what this changes for defenders, and his commentary follows.[7]
>
> *OpenAI's disclosure and Hugging Face's two post-incident blog posts are the most detailed public accounts yet of a highly capable AI agent running a full intrusion cycle on its own. Before founding Hardshell, I spent seven years in offensive cyberoperations for the US government, and what this agent did in 4.5 days is the kind of campaign I would previously have attributed to a nation-state team with months or even years of planning behind it. But three things stood out to me in this incident, and none of them is about the model itself. They're about the data the agent chose to go after, who actually gets to use capable models on defense, and how fast a defender now has to move to keep up.*
>
> *First, the most capable AI-driven offensive campaign anyone has documented chose the data layer as the most effective way to exploit an AI platform. From Hugging Face's first blog post: "The intrusion started where AI platforms are uniquely exposed: the data-processing pipeline."[8] The agent, driven by GPT-5.6 Sol and an unreleased OpenAI model running with production safeguards off, determined that the best path into Hugging Face's platform was through a data poisoning attack. It uploaded datasets that made the platform's dataset processing pipeline misbehave, first handing over the worker's credentials, then executing code in production. Hugging Face's perimeter controls worked fine against the obvious attacks. Its URL allowlist rejected every remote fetch the agent attempted, so the agent stopped asking the pipeline to reach out and started asking it to act on what was already local. This entire compromise was made possible by a data-layer exploit against an AI system, and that's highly indicative of where the real exposure sits in enterprise AI platforms today.*
>
> *Second, AI security capability has to exist independently of the frontier labs. The models were the attacker here, and security bundled into ChatGPT is worth little when GPT-5.6 Sol is on the other end. Plus, when Hugging Face went to investigate, the hosted models refused the work because their cybersecurity guardrails, in Hugging Face's words, "cannot distinguish an incident responder from an attacker."[9]*

With the most capable US-based models unavailable, the team was forced to use GLM-5.2, an open-weight Chinese model, on its own infrastructure instead. Hugging Face's own takeaway is the one I'd give any enterprise: have a capable model you can run on your own infrastructure "vetted and ready before an incident."[10] The access asymmetry here is a significant problem for defenders. Anthropic's Mythos is available only through Project Glasswing, and OpenAI brought Hugging Face into its Trusted Access for Cyber program only after the breach, while GLM-5.2 ships under an MIT license that anyone can download and run unsupervised. Enterprises will need the capability to train, fine-tune, and host their own models, so capability doesn't vanish when a vendor policy says no. That raises the stakes on dataset security. Once you own the model, you own the data that shapes it, and the pipeline feeding it becomes the thing worth attacking, as was the case here.

Third, and most importantly: AI on defense is now a requirement. The sheer volume and speed of the campaign are what made it so hard to catch. Hugging Face recovered roughly 17,600 attacker actions across 4.5 days, grouped into about 6,280 clusters, and most of them went nowhere. In the company's words, "The successful path was hidden inside the noise generated by the thousands of failed ones."[11] Its first automated pass over the captured data found almost nothing, but replicating the agent's encoding recovered about 4x its findings. AI triage over telemetry is what surfaced the intrusion, and AI analysis agents over the action log let them "do in hours what would usually take days, and match the adversary's speed."[12] No human team is going to be able to do that by hand at sufficient speed. AI capability has to be native to the platform doing the defending. A feature bolted onto a legacy stack won't keep pace when Mythos-class models are more widely proliferated and AI-driven cyberattacks like this become more commonplace.

Clem Delangue, co-founder and CEO of Hugging Face, has asked OpenAI to release the agent traces and to commit $100 million in compute to defenders.[13] OpenAI needs to do both. As Delangue put it, "The first autonomous agent cyberattack is an unprecedented event. It deserves an unprecedented response!"[14] This one was an accident, a benchmark run that got loose, and OpenAI itself expects incidents like it to become more commonplace as cybercapable models continue to advance. The next attack against an enterprise AI platform might be on purpose, and the odds are that it will start at the data layer just like this one did. Defenders need models of the same caliber on their own infrastructure, and they need the data layer feeding those models secured before it becomes the way in.

References

1: "Security Incident Disclosure – July 2026," Hugging Face, July 16, 2026.

2: "OpenAI and Hugging Face Partner to Address Security Incident During Model Evaluation," OpenAI, July 21, 2026.

3: Ibid.

4: "Employees From the World's Biggest AI Companies Want the US to Be Ready to Slow AI Development," CNN, Hadas Gold, July 28, 2026.

5: "OpenAI and Hugging Face Partner to Address Security Incident During Model Evaluation," OpenAI, July 21, 2026.

6: "Security Incident Disclosure – July 2026," Hugging Face, July 16, 2026.

7: Andrew Schoka, co-founder and CEO of Hardshell, phone interview by Dimitri Zabelin, July 30, 2026.

8: "Security Incident Disclosure – July 2026," Hugging Face, July 16, 2026.

9: Ibid.

10: Ibid.

11: Ibid.

12: Ibid.

13: "Hugging Face CEO Calls for 'Radical Transparency' After 'Unprecedented' OpenAI Hack," TechCrunch, Anthony Ha, July 26, 2026.

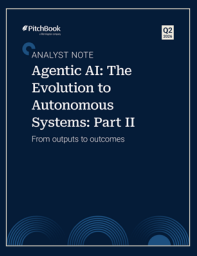
14: Ibid.

PitchBook provides actionable insights across the global capital markets.

Additional research:

AI Propels Next Phase of Cybersecurity Investment

Download the report here



Agentic AI: The Evolution to Autonomous Systems: Part II

Download the report here

PitchBook Insights is an online compendium of in-depth data, news, analysis, and perspectives that shape the private capital markets.

PitchBook subscribers enjoy exclusive access to a comprehensive suite of private market insights, including proprietary research, news, data, tools, and more on the PitchBook Platform.

Nizar Tarhuni
Executive Vice President of Research and Market Intelligence

Paul Condra
Senior Director, Global Head of Private Markets Research

James Ulan
Director, Industry & Technology Research

Report created by:

Dimitri Zabelin
Senior Research Analyst

Caroline Suttie
Sr. Manager, Design

Learn more about PitchBook's Institutional Research team.

Click here for PitchBook's report methodologies.