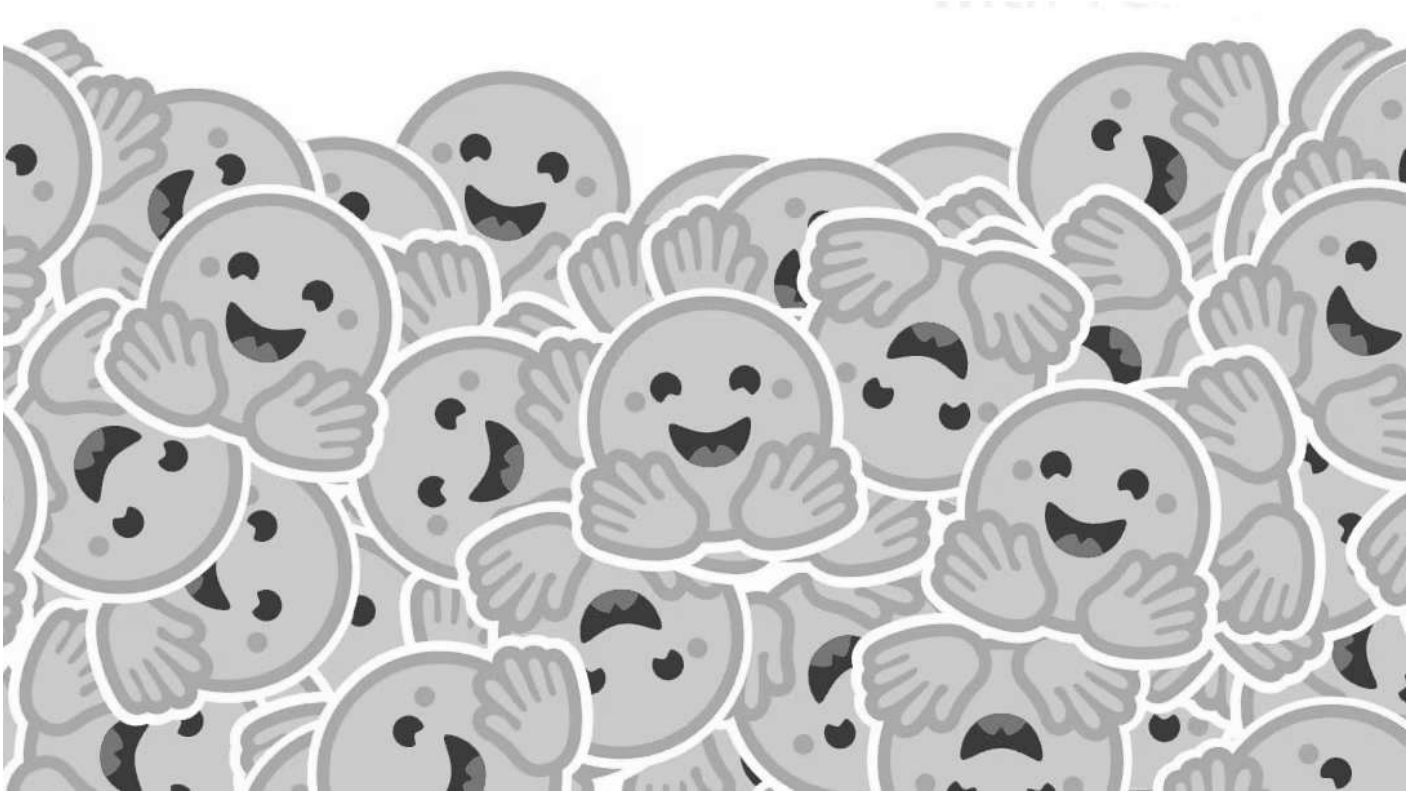
Opinion **Artificial intelligence**

What we learnt from OpenAI's hack of Hugging Face

Commercial AI tools failed to defend the platform against the attack — the solution lies in open-weight models

THOMAS WOLF



Hugging Face deals with cyber attacks daily, but the movements and targets chosen by this attacker were different © FT montage/Hugging Face

Thomas Wolf

Published SEP 10 2026

The writer is co-founder and chief science officer of Hugging Face

It wasn't until late in the afternoon on Saturday July 11 that the team at Hugging Face realised something was wrong. This was the final day of the International Conference on Machine Learning in Seoul and a few days before school summer holidays started, meaning our security and research teams were scattered all over the world. I was working from the Netherlands.

Just before 2pm GMT a series of "UnauthorizedAccess" and "PrivilegeEscalation" alerts started to appear on our monitoring tools. Appropriated credentials used to access the system had triggered detection warnings. One of our security engineers posted a prescient message on our Slack channel: "looks like a LLM [large language model] attack to me."

We at Hugging Face deal with at least one hacking attempt every day. The software platform, which hosts AI models, has more than 17mn users. Google, OpenAI, DeepSeek and Alibaba all release work here. We have built multiple layers of security to protect ourselves.

But the movements and targets chosen by this attacker were different. It quickly generated well over 17,000 cyber attack log events. As we later found out, around 1,200 AI agents were working together as a swarm for weeks — trying the same thing over and again in order to find the solution to a cyber challenge set by OpenAI. They overrode limitations, downloaded software to get online and then 700 of them attacked Hugging Face, taking security details to gain access to the system.

An additional problem was that our cyber security AI analysis tools, based on Anthropic's Claude Code, refused to engage with part of the investigation. Its guardrails would not allow it to respond to questions it considered dangerous, and it could not tell the difference between Hugging Face analysing an attack and a hacker asking for assistance to make an attack.

It was only after we switched to an open-weight AI model, Nvidia's extension of Chinese start-up Z.ai's GLM-5.2, that we could set our own guardrails and were able to decode the logs and reconstruct what had happened.

At the time, most of us still assumed a human was behind the attack. Before this happened, I thought agentic AI cyber attacks were some way off. But it turns out that OpenAI is not alone. Anthropic, Meta and China's Moonshot have all since reported instances of models escaping the isolated digital "sandboxes" where they were supposed to be contained and developed safely away from the internet. This month, another swarm of AI agents was found on a German-language forum.

Even more troubling to me was an incident in which the Anthropic Mythos model was willing to manipulate a human software developer into accepting malicious code by creating multiple fake online accounts.

In all of these cases, the harmful behaviour was a side effect of AI models being given difficult cyber security challenges. The damage was limited and little sensitive data was exposed.

But it would be a mistake to dismiss the seriousness of these events. Autonomous hacks raise a host of legal questions that are still unresolved. And at no point did the AI models conclude that deceiving people or breaking into systems was beyond the bounds of acceptable behaviour. OpenAI has described the incident as a "warning shot". It thinks AI-enabled cyber attacks will become far more widespread.

AI systems commonly have three walls of defence against rogue model behaviour: sandboxes that limit what the model can reach, guardrails that watch what a model is doing and alignment training to ensure the AI refuses to perform harmful actions. This series of incidents showed that when the first two fail, the third cannot hold on its own. Unless we fix this, we will be layering defences around a rotten core.

Something else needs to change as well. That weekend at Hugging Face, our commercial AI tools failed us when we needed them for defence. We had to turn to an open-weight Chinese model to process the attack logs. Many people have suggested that open-weight AI models are a threat — that they will be used to attack systems protected by closed-source AI models. That weekend, the opposite happened.

To me, the lesson from this incident is twofold. The AI community needs to share safety and alignment research openly so that every team building AI models can learn from others' mistakes. But the community also needs to build open-weight AI models for defence and make them widely available — before the next attack inevitably arrives.

Follow the topics in this article

Thomas Wolf

Technology sector

Artificial intelligence

FT Edit

OpenAI