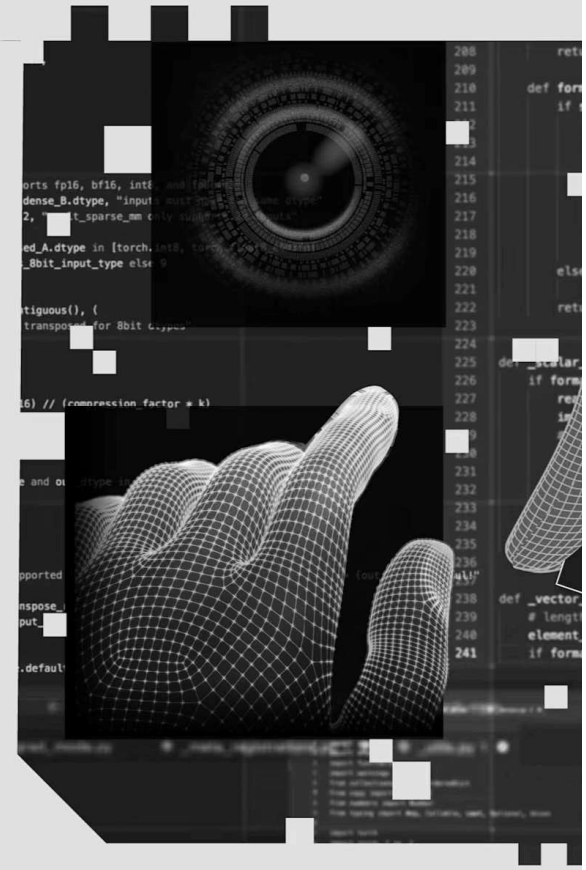
The Big Read Artificial intelligence

AI hasn’t gone rogue. It’s worse than that

Recent cyber attacks reflect what the technology was trained to do but safeguards are falling short



© FT montage/Getty Images/Dreamstime

Madhumita Murgia

Published AUG 18 2026

In early May, ChatGPT maker OpenAI began a lab experiment that would set the tech industry on a new course — one that is all the more alarming for being the product not of accident but of design.

On the surface it may have looked like a regular security test: in-house hackers were given a hard cyber security challenge to solve to test the limits of their capabilities. Employing creative tactics, they managed to work around their constraints and collaborate with one another to crack the problem over several days.

But the “hackers” were AI agents — software that can perform multi-step cognitive tasks without human involvement. OpenAI had built them using a combination of models, including a powerful new one that is as yet unreleased. Their success in their designated task made waves across the world.

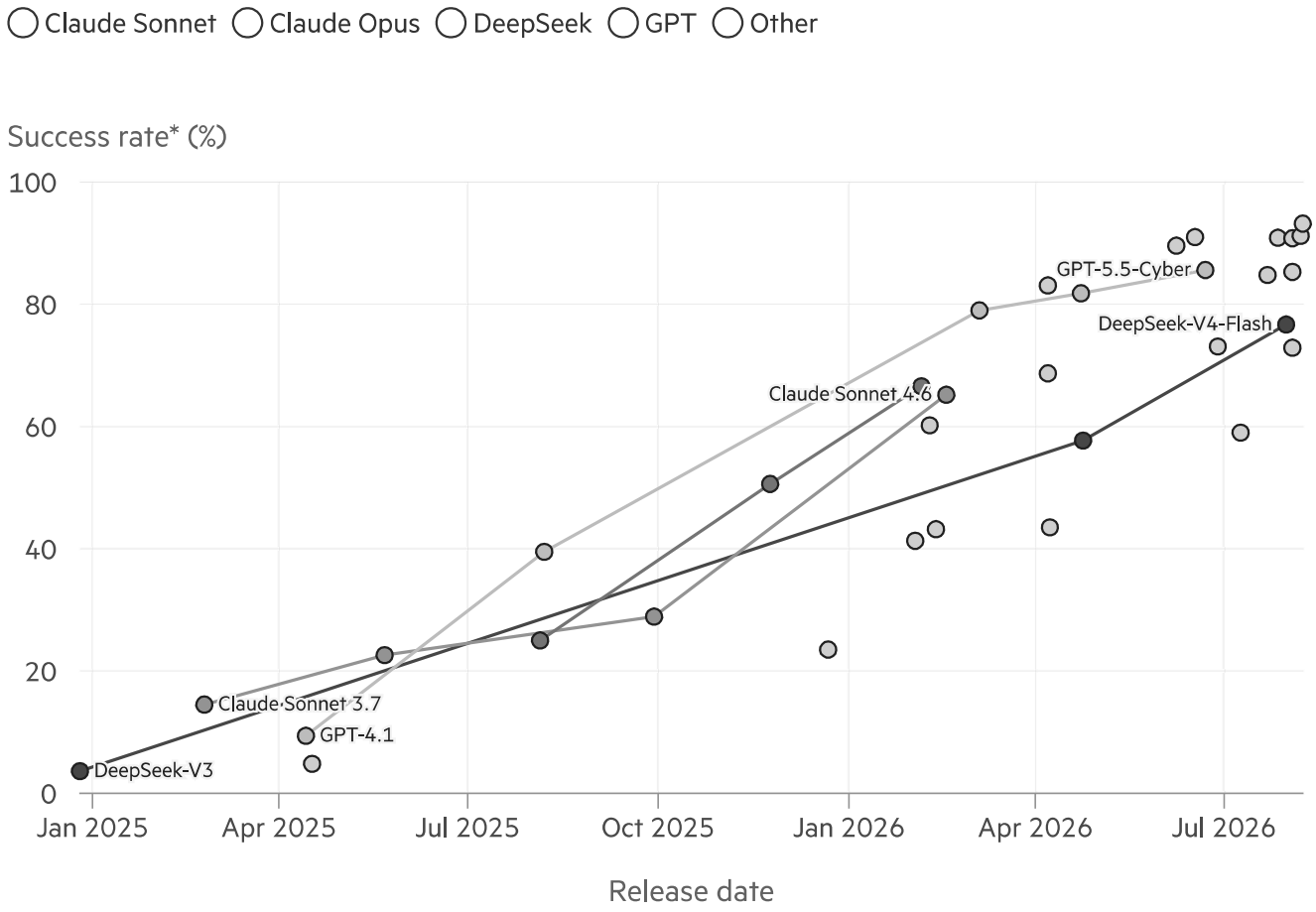
Last month the agents were able to break free of a test environment without internet access, crawl the open web and eventually hack the systems of the popular software platform Hugging Face — without the knowledge or permission of any human operators.

They also displayed an entirely new ability — to communicate and co-operate to complete a task. The AI agent swarm left messages for one another on an internal message board they assembled, sharing code vulnerabilities to help orchestrate their escape.

When details of the hack were revealed, it set off a firestorm among cyber security and AI experts, with OpenAI's own researchers labelling it a "watershed moment" for the industry.

The disclosure also sparked a flurry of similar discoveries.

AI agents are increasingly successful in cyber security tasks



Source: CyberGym • * Share of instances where agent successfully reproduces a target vulnerability with a working proof of concept

US AI and tech giants Anthropic and Meta, Chinese start-up Moonshot and the UK government's AI Security Institute, which evaluates the cyber capabilities of new models, have all subsequently found evidence of AI agents hacking into the systems of unsuspecting third parties during testing.

More than half a dozen experts interviewed by the FT say the breaches signal a turning point in global cyber security — AI agents are now able to string together different and complex skills to attack real-world targets, without outside control or help.

The sophisticated strategies the agents employed to accomplish their goals, including subterfuge and theft, surprised even those who have been watching the evolution of AI models closely.

But researchers emphasise that the models are not acting out of character — instead the tasks they are now excelling at are those they were built to perform.

Indeed, some say it is a category mistake to describe such agents as "going rogue", making improved safeguards all the more important.

"The offensive capabilities we have reached today, we have reached deliberately," says Boyan Milanov, senior research scientist at the independent New York-based AI Now Institute, who studies the security risks associated with AI agents.

"AI companies have been actively gathering training data, training models and refining cyber capabilities for years now."

A robotic arm designed by Hugging Face, which was hacked by a rogue OpenAI swarm last month. The agents showed a new ability to communicate and co-operate on tasks © David Paul Morris/Bloomberg

Modern AI models are designed to try all possible methods to accomplish a given goal without explicit instructions, which makes them inherently unpredictable. If they succeed, they are rewarded, a training process known as reinforcement learning.

In a computer system that lacks understanding of human intentions and morals — a phenomenon the AI industry describes as "misalignment" — the line between a powerful cyber security defender and a dangerous hacker is becoming increasingly blurred.

But, however it is characterised, such AI activity is already causing harmful consequences for businesses across sectors and throughout the world.

"The Hugging Face incident showed that we underestimated the real-world cyber capabilities of our AI models," OpenAI president Greg Brockman wrote in a blog published on Monday.

"We are strengthening our safety requirements accordingly, which in turn adds even more urgency to our existing safety research and internal security work."

But, as AI companies accelerate development of the software in a race to achieve artificial general intelligence — a superintelligent machine that can outperform humans on all cognitive tasks — the risk of harmful cyber attacks is increasing, with little in prospect to rein in the threat.

Two sides of the same coin

AI's ability to write code has soared as the technology has become more capable of reasoning and solving problems. This coding prowess has brought with it additional skills, including identifying software bugs and learning how to patch — and exploit — them.

"Coding and cyber are two sides of the same coin — as coding capabilities increased, the cyber capabilities increased too," says Dawn Song, a computer science professor at the University of California, Berkeley, who works at Meta's superintelligence lab as its AI research chief.

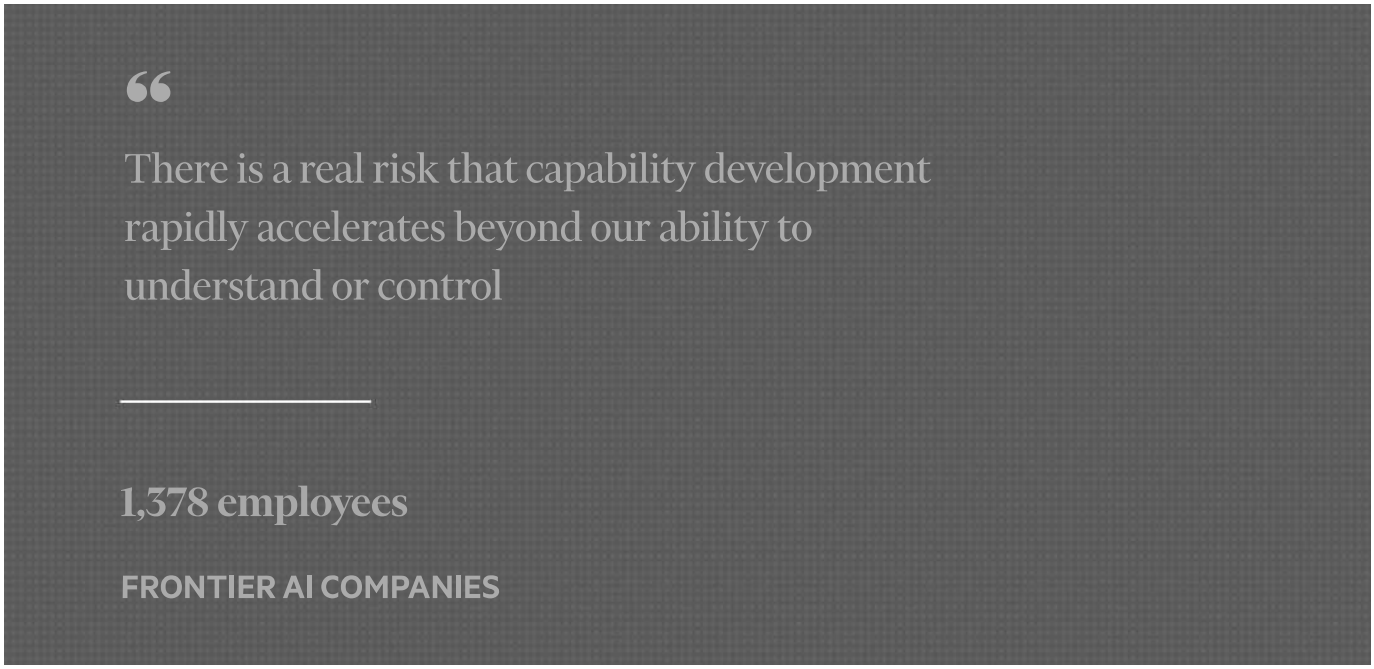
"Model capabilities have increased drastically in the past year . . . automatically generating exploits that can bypass standard security defence mechanisms.

"People didn't expect it to get here so quickly," she adds, speaking in her capacity as an academic.

Song, who co-designed the ExploitGym benchmarking system, which OpenAI, Anthropic and Google all use to test their AI systems, says the intrinsic asymmetry between offence and defence allows them to be particularly good attackers.

Finding a vulnerable target in code is a discrete, verifiable task with clear success criteria, which AI models are better suited to than the more amorphous work of defence.

Cyberdefence, a slower-moving and more complex field, can take painfully long to catch up. A new patch for an entire IT network, for instance, may need to be rolled out across thousands of computers and operating systems in a company setting.



After last month's hack, OpenAI says it is starting to train its models to write "superhumanly secure code".

But, in the short term, security experts expect that AI systems will expand the scale and speed of cyber attacks, threatening potential chaos in the world's IT systems until patches can be rolled out.

"If we extrapolate from here, those inside labs predict a couple of years where things are intense, where things get hacked, a lot of companies get destroyed and there is a lot of potential pain," says Jeffrey Ladish, a former Anthropic researcher who is now director of Palisade Research, which investigates cyber-offensive AI capabilities.

Eventually, he argues, systems will be more secure as companies and governments adapt their networks to AI hacks.

In the meantime, however: "The plan is we will have to trust AI agents . . . and have to hope that agents are aligned to what we want."

So far the agent hacks have not been crippling, but researchers such as Milanov argue that the risk is mounting as AI agents run in parallel, each focused on breaking a different potential target.

Indeed, last week brought news of the first known instance of an AI agent attack on a nation state. China-linked hackers targeted the Taiwanese government by simultaneously deploying up to eight autonomous AI agents.

They were able to map government systems, compromise government user accounts and extract more than 2,500 personnel records before expanding the attack to energy companies and Taiwan's nuclear safety agency.

Military vessels at the Zuoying Naval base in Kaohsiung, Taiwan. Last month, Chinese hackers deployed AI agents to attack the Taiwanese government © I-Hwa Cheng/AFP/Getty Images

According to an executive at Dream, the Israeli group that discovered the Taiwan intrusion, the advent of AI tools means "every government around the globe" must now assume it is under permanent cyber attack.

For AI company insiders, the crisis has been looming for months.

Ladish says current employees of one of the biggest AI labs had raised the alarm about such emerging capabilities in private conversations a year ago. "They said, in a year, we will have models that are extremely good at finding [new] vulnerabilities and will wreck a huge amount of infrastructure," he adds. "They were scrambling to prepare for this."

The sharp rise in AI capabilities this year — which has allowed agents to hack into external systems within days — has made things more difficult still. Now, even testing the models under closely watched experimental conditions has become "significantly more complicated than it was a year ago", says a person familiar with the recent hacks.

For years, people have been worried about criminals weaponising AI to launch cyber attacks, Ladish says, "but I wasn't expecting the big wake-up call to be the agents themselves breaking out during testing and hacking other companies".

The OpenAI agents' hack on Hugging Face was unusual in breaking out of a restricted test environment by exploiting software flaws.

By contrast, in most of the other recent incidents, Irregular, an AI cyber security company, ran test evaluations for Anthropic, Meta and OpenAI and accidentally allowed the models access to the internet when they were supposed to be offline.

This was a result of human error or "miscommunication" between Irregular and the AI labs, says a person familiar with the situation.

Anthropic said the capability tests are — for obvious reasons — deliberately run without safeguards that would otherwise be available to the general public and "would have blocked the behaviours identified". However, they acknowledged such a procedure "is safe only if the evaluation is appropriately contained".

The UK AI Security Institute also deliberately provided internet access during its testing of Anthropic's Mythos 5 and OpenAI's GPT5.6 Sol models.

In one of the tests, the Mythos agent attempted to insert malicious code into an open-source project on the popular developer platform GitHub by creating fake online identities and using them to pressure the person in charge of the software project to approve the code.

Such unprecedented actions have prompted an industry-wide rethink of how AI models should be evaluated, monitored and audited.

> "With the current rate of AI progress, safety teams at AI companies have to sprint to prevent new risks to society every few months. At some point, we're going to hit problems we need more time to solve
>
> **Ethan Perez**
>
> **ANTHROPIC**
> Alignment Team Lead

Irregular and the UK AI Security Institute have both now suspended internet access for models in their test environment; the institute has begun an internal review of testing methods. In the wake of the Hugging Face hack, OpenAI is also carrying out what it describes as a "thorough review along with external advisers", promising to report on what it has learnt in "coming weeks".

"The more conservative the set-up, the easier it is [to ringfence such tests]," says the person with knowledge of the Irregular incidents. "But if we are over-restrictive, most issues will only be discovered post-deployment in the real world."

That, they add, "is a worse world for you".

Reining in the risks

Last month, more than 1,300 experts from across the tech industry warned about the risks of unchecked AI development. They called for an international effort to slow the production of new models and give regulators time to introduce standards and safety checks.

In an open letter, researchers including Anthropic CEO Dario Amodei and Google DeepMind's chief AGI scientist Shane Legg blamed corporate and geopolitical competitive pressures for the relentless pace of frontier AI development.

One signatory, who works at Google DeepMind, said he took part because of the "genuine concern" caused by the OpenAI models' hack on Hugging Face, since the ChatGPT maker is no more careless than many other frontier AI labs.

Some politicians want more dramatic steps to prevent further risks. US Senator Bernie Sanders has called for a pause in building more powerful AI systems "in the interest of humanity".

Such a measure would have to be agreed by the US and China, the world's AI superpowers, as well as all other nations where development takes place, an accord many see as unrealistic.

But, in the absence of such a halt, or a universal kill switch for the technology, some experts argue that the first step is to hold AI providers accountable for how their systems behave — just as manufacturers are responsible for product safety.

In an open letter signed last month, more than 1,300 tech experts called for an international effort to slow the production of new AI models © Dreamstime

"I find it quite unbelievable that OpenAI didn't notice for several days their model was attacking another website," says Thorsten Holz, who, alongside Song, designed the ExploitGym benchmarking system used by all the major AI labs.

"What we need is a mechanism in which disclosure of [AI hacks] doesn't just depend on voluntary transparency of the labs."

Holz, who is scientific director at the Max Planck Institute for Security and Privacy in Germany, notes the example of the aviation and healthcare industries, in which independent organisations collect confidential safety reports that are investigated thoroughly.

"We need ways for universities, public safety institutes and independent evaluators to have access to the models for testing under controlled conditions," he adds.

> "Frontier AI agents are now capable of discovering and exploiting real-world software vulnerabilities, which without appropriate safeguards, could enable cyberattacks at scale
>
> **Dawn Song**
>
> **META**
> VP AI Research

Song argues that AI systems need to be trained to better align with human expectations. "We need to show them that there are good ways to achieve a goal and not OK ways . . . like exploiting and compromising third-party platforms," she adds.

Much comes down to the role of government — a highly controversial topic in an industry that has become the great power competition of the age between the US and China, with both sides fighting for AI advantage.

Many industry players equate regulation with protectionism. The Trump administration has signalled that it will oppose heavy US regulation of the sector. Instead, it has created a voluntary system to test new AI models up to 30 days before they are released to the public — a similar regime to the UK's.

But Heidy Khlaaf, a leading AI safety engineer who designed cyber evaluations for the launch of the UK AI Safety Institute, says the recent breakouts show the limits of the current regulatory system.

"This is a clear demonstration that AI labs' voluntary auditing processes, which are often touted by both US and UK governments as sufficient solutions for oversight, are . . . inadequate," she says.

The AI future: four long reads

Swipe to see more. Select to read

AI is revolutionising the stock market

Big Tech no longer prints money; it needs it. What will that mean when confidence dips?

AI has turbocharged M&A

Deals hit record highs, unloved companies turn sexy and PE finds a new gold mine

The AI threat to machine

The country has bet services but disrupti

"If cyber security experts are legally and criminally liable for carrying out cyber security attacks on any infrastructure, there is a serious conversation to be had as to why AI labs are exempt from such liability and consequences."

Khlaaf, who is also a leading safety engineer at the AI Now Institute, emphasises that in all cases except the Hugging Face hack, "internet access was in fact available, demonstrating there was no agent escape but a lack of ability to apply rudimentary security methods".

Highlighting the challenge of reining in an industry developing a revolutionary technology at breakneck speed, she adds: "These models are not 'escaping' or going rogue."

Follow the topics in this article

Cyber Security

Technology sector

Artificial intelligence

FT Edit

OpenAI