AI arms race in line for a reckoning after OpenAI hacking incident

OpenAI has conducted this type of model testing for years, and there have been early warning signs in previous models of systems that will act maliciously and attempt to escape environments.

In April, Anthropic's Mythos model also gained internet access and published details of a security exploit online publicly, beyond what researchers anticipated the model would do.

Mythos, and Anthropic's subsequent Fable model, made reverberations in the cyber security community and caused governments around the world to home in on the idea that attacks on digital and critical infrastructure will be increasingly AI-led and autonomous.

Jake Moore, global cyber security adviser at ESET, a cyber security company, said OpenAI would inevitably use the breach as a marketing tool, given how much rival AI developer Anthropic benefited earlier this year from similar concerns. "I just don't think that OpenAI had a matching story and so maybe they'd been waiting for something like this," he added.

Following this incident, many in the AI safety and cybersecurity communities have called for regulation or standards to avoid a repeat. Altman is expected to brief White House officials next week on the next generation of AI systems.

As systems move towards more autonomous capabilities, less

desirable behaviors, such as hacking or disobeying instructions, may emerge. Hobbhahn, of Apollo Research, said that in order for agents to become effective, they have to work unsupervised for long periods. “They have to have more agency; there’s just no way around it.”

He added: “People say, ‘It’s just a tool, it does what you wanted it to do and nothing else and it just follows exactly your intention and instructions.’ And I think people should be really prepared for agents having their own goals, acting autonomously for days, and those goals not necessarily being aligned with yours.”

Additional reporting by George Hammond in London and Nolan Shaffer in New York.

© 2026 The Financial Times Ltd. All rights reserved. Not to be redistributed, copied, or modified in any way.

https://arstechnica.com/ai/2026/07/ai-arms-race-in-line-for-a-reckoning-after-openai-hacking-incident/