**Columnists**

Friday 24 April 2026

# Anthropic's Mythos AI heralds a new era of cybersecurity warfare

**The company's latest large language model has found chinks in the armour of 'every major operating system and web browser'. Now the challenge is to stop it falling into the wrong hands**

John Naughton
*Columnist*

Share

Mythos, says Collins dictionary, is "the complex of beliefs, values, attitudes, etc, characteristic of a specific group or society". It is also the name Dario Amodei's Anthropic has given to its latest AI model – officially Claude Mythos Preview – which has arrived as a bombshell, and whose reverberations we will soon begin to notice.

How come? It turns out that Mythos, which is a general-purpose large language model (LLM), is also a world-class hacker. Or, to put it less luridly, it is striking for its ability to discover security vulnerabilities in networked systems and exploit them. Anthropic said that it has already found thousands of high-severity vulnerabilities, including some "*in every major operating system and web browser*". It added: "Given the rate of AI progress, it will not be long before such capabilities proliferate,

potentially beyond actors who are committed to deploying them safely." It also said: "AI have reached a level of coding capability where they can surpass all but the most skilled humans at finding and exploiting software vulnerabilities."

Discovery of this capability persuaded Anthropic that it would be unsafe to launch Mythos on an unsuspecting world. Accordingly, the company decided on Project Glasswing; the codename for making the model available only to a number of large – and presumably trusted – corporate bodies such as Amazon Web Services, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorgan Chase, the Linux Foundation, Microsoft, Nvidia and Palo Alto Networks, plus some government agencies such the UK's AI Security Institute (AISI). This is so Anthropic could evaluate the dimensions of the security threat posed by Mythos and harden its own systems.

On 13 April, AISI's analysis of Mythos's capabilities largely supported Anthropic's assessment. Of the LLMs tested, it was the only one that, during a simulation, accomplished a full network takeover; a 32-step corporate network attack spanning initial reconnaissance through to full network takeover, which the AISI estimated would take the most skilled human 20 hours to complete.

When Anthropic announced Mythos was too dangerous to release, cynics saw it as just an attempt to ramp up the hype around a new product. After all, hadn't OpenAI tried that stunt with GPT-2 So was Anthropic engaging in hype inflation or being a practitioner of responsible AI?

Answer: probably a bit of both. On the responsible AI side, the company's caution about the public release of the model induced serious folk such as the chair of the US Federal Reserve, Jerome Powell, and the US treasury secretary, Scott Bessent, to summon the chief executives of big American banks for an emergency meeting about the dangers of Mythos.

On the hype side, Anthropic's boss suddenly found himself invited to the White House for a meeting with Donald Trump's chief of staff, Susie Wiles, apparently to discuss the use of Mythos within the US government. And then it was revealed the National Security Agency was *already* using Mythos, presumably to hunt down zero-day vulnerabilities (unknown security flaws that need immediate fixes to thwart hackers) in its own – and other's – systems.

Pause to ponder that for a moment. Up until then, Anthropic had been condemned by the administration as an "unacceptable" national security risk, and now we find the erstwhile pariah turning up at the heart of the deep state. Like much else in Trumpland, you couldn't make it up.

**Related articles:**

Confidence crisis in software sector delays Visma IPO

AI cannot be allowed to trample on our mental health

These ironies shouldn't distract us, though, from the realisation something momentous has happened. We've always known that there is no such thing as a secure networked computer, but until comparatively recently, hacking has generally been a pretty arcane craft requiring significant technical skills. The overall trajectory over four decades is striking: from lone hobbyists, jokers and phone "phreakers" in the 1980s, through organised criminal networks in the 2000s, to state actors treating "'cyberspace as a domain of warfare" by the 2010s.

The arrival of Mythos – and the models that will follow it – heralds a new cybersecurity era. In it, AI will, on the one hand, enable us to detect hidden vulnerabilities of which we were hitherto unaware; and on the other, enable bad actors to find and exploit those vulnerabilities, while our role is to be confused spectators of the ensuing battles in cyberspace. And maybe that kind of warfare is closer than we think; even as I write, Bloomberg is reporting that Anthropic's model is already being accessed by "unauthorised users".

So perhaps it's appropriate that the term "mythos" was coined by the ancient Greeks, for whom it was the language of narrative, tradition and imaginative storytelling. And could it be that the subliminal tale Anthropic's new creation is narrating involves a species that bet its collective future on a technology that turned out to be much more fragile than it ever knew?

## What I'm reading

**Challenged universities**

You Are Not a Function is a thoughtful essay by Brendan McCord about the trap into which universities are falling.

**Hot topic**

A sobering blogpost by Mike Brock about phase transition in democracy is The Boiling Point.

**Must see**

Notes on Hong Kong is Rohit Krishnan's guide to that amazing place.

*Photograph by Krisztian Bocsi/Bloomberg via Getty Images*

---

Artificial intelligence

---

FAQs
Careers
Shop
About us
Editorial policy

## Follow

The Observer

The Observer Magazine

The ObserverNew Review

The Observer Food Monthly