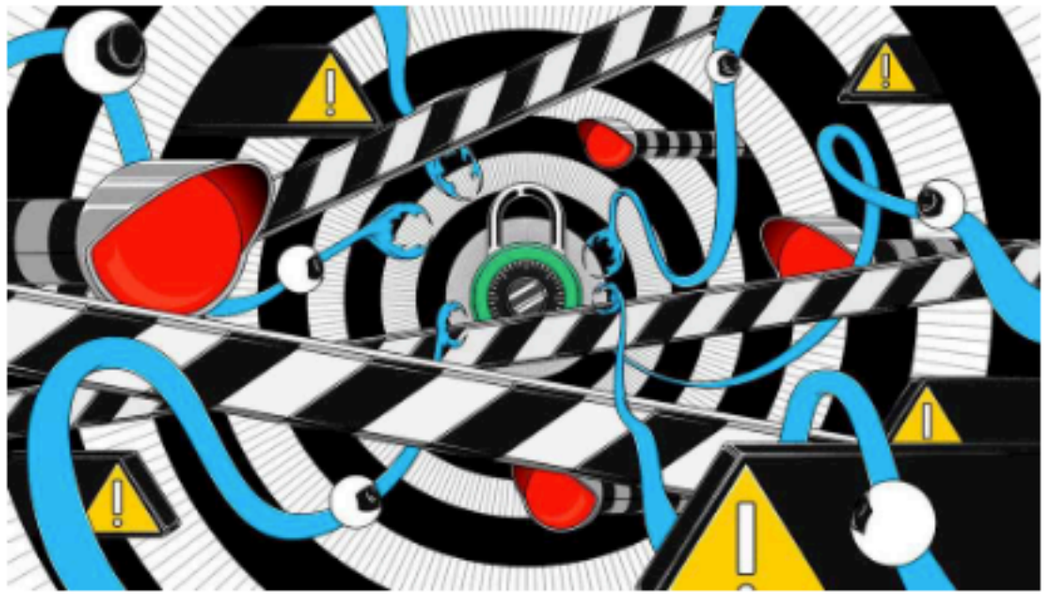
Science & technology

Science & technology | Examining the Mythos

How AI hackers will shake up cyber-security

The technology could eventually favour the defenders—but expect a bumpy ride

April 16th 2026



TECH FIRMS usually create buzz around products they plan to release. Anthropic, an American artificial-intelligence lab, has managed to create excitement—and a good deal of worry—around something it plans not to. On April 7th the firm announced that a new AI model it had developed,

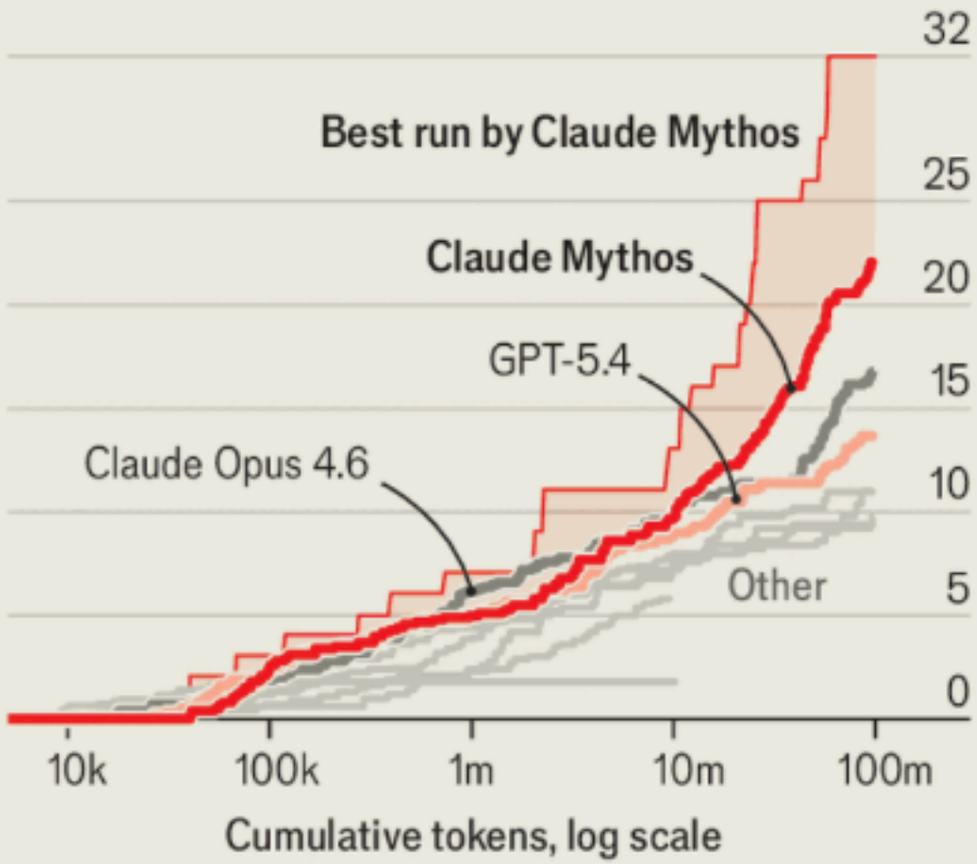
dubbed Mythos, would not be released to the general public. Instead, under an initiative called Project Glasswing, whose 12 founder members include Apple, Google and Nvidia, access would be strictly controlled.

The problem is not that Mythos is buggy or unreliable. Allegedly, it is that it works so well that releasing it would put the world's digital infrastructure at risk. According to Anthropic, the model has surpassed "all but the most skilled humans" when it comes to finding and exploiting security holes in everything from popular operating systems to the cryptographic software that secures e-commerce and financial networks. And it can find those vulnerabilities with only the bare minimum of human help. Not to be outdone, a few days later OpenAI, one of Anthropic's competitors, announced a closed version of its own hacking-friendly model, named GPT 5.4 Cyber.

A world of "vibe hacking", in which amateurs can use AI models to find flaws in software—and perhaps even write the "exploits" needed to crash them, hold them to ransom or even take control of them remotely—sounds terrifying. Shortly after Anthropic's announcement Scott Bessent, America's treasury secretary, hosted a meeting of bank bosses to discuss what AI-enabled hacking might mean for their businesses. Financial regulators in Britain organised a similar meeting a few days later. But security researchers themselves seem guardedly optimistic. "In the medium term I think this will be a mess," says Bruce Schneier, an American computer-security expert. "But in the long run I think it will actually be good for the defenders."

Plan of a hack

AI-model performance in cyber-security challenge, 2026, average steps completed out of 32



Source: AI Security Institute

Since Anthropic has released only limited information about Mythos, the degree to which the new model really is revolutionary rather than evolutionary is hard to judge (what might politely be termed a "vigorous debate" is raging online). Testing by the AI Security Institute, a British government agency, found that Mythos was neck-and-neck with other models on relatively simple cyber-security tests, but noticeably ahead in a more advanced one that requires a model to complete dozens of steps before successfully taking over a target machine (see chart).

The chief thing Anthropic's researchers investigated was Mythos's ability to unearth bugs that hackers could use to attack or gain control of other computers. They looked specifically for bugs that had never been found before (known as "zero-days" in the jargon). Finding those would prove the model was doing novel work, and not simply regurgitating known bugs it had come across in its training data.

Zero-days lurk everywhere, says Jeff Williams, a co-founder of Contrast Security, a software firm, and of the Open Worldwide Application Security Project Foundation, a non-profit dedicated to improving the security of software. Although Mythos is said to have found "thousands" of high- or critical-severity flaws, Anthropic is keeping most secret until they can be fixed. But the firm did reveal details of some, including one in FreeBSD, a widely used operating system, another in FFmpeg, a video-and-audio code library, and a

third—which remains unfixed—in software vital to cloud computing.

Many of the bugs reported by Anthropic are, if not simple, then at least comprehensible. They are the sorts of things a human could plausibly have found. They seem to be the sort of thing other AI models could have found, too. In a blog post published shortly after Anthropic's announcement, Stanislav Fort, a founder of AISLE, an AI-focused cyber-security company, described using several smaller, older models to find the same bug in FreeBSD. Citing his own firm's experience with AI-powered bug-hunting, Dr Fort reckons the AI cyber-security frontier is "jagged", with no model having a clear edge.

Everyone agrees that the state of the art is advancing quickly. Until recently AI bug-hunting was prone to generating false positives or trivial results. "One change I've noticed in the past couple of months is that a lot of these AI-generated bug reports are increasingly of good quality," says Mr Schneier. An update in January to OpenSSL, which helps ensure secure connections between websites, fixed a dozen security flaws found by AI models employed by Dr Fort's firm. In March Anthropic itself announced that an older, pre-Mythos version of Claude had found almost a fifth of all the high-severity bugs fixed in Firefox, a web browser, in 2025.

As the growing power of AI models makes finding bugs easier, says Mr Schneier, the question becomes whether attackers can exploit them more quickly than defenders can fix them. This is where Project Glasswing comes in. Anthropic says it is expanding Glasswing to another 40 digital-infrastructure organisations, so they can use Mythos to harden the software on which the internet depends. Anthropic hopes that giving them access now, before similarly powerful models become widely available, will leave them time to find and fix as many bugs as possible.

All the researchers The Economist spoke to thought that, in the long run, AI-enabled hacking would probably help defenders more than attackers, by allowing companies to more thoroughly check their software before it is published. But there is plenty of short term to worry about. For one thing, AI checking is not cheap: Anthropic says one of the bugs it found cost the AI lab nearly $20,000-worth of tokens to find. For software such as Linux, a family of widely used operating systems which are at least partly maintained by vol-

unteers, that would be a steep price. And much of the code out in the world—running on home routers, smart gadgets like TVs or fridges and industrial machinery—has nobody maintaining it at all. In such cases, attackers could have a field day. ■

Curious about the world? To enjoy our mind-expanding science coverage, sign up to Simply Science, our weekly subscriber-only newsletter.

Science & technology | Serving them right

How to make buffet breakfasts less wasteful

A computer model has found some counterintuitive solutions

April 16th 2026

BREAKFAST IS THE most important meal of the day, and how it is served matters, too. Take the classic hotel buffet breakfast. Or, maybe, don't: when people do, they take much more than they eat. Compared with ordering from the menu, all-you-can-eat breakfasts waste more food—up to