AI has crossed a threshold – what Claude Mythos means for the future of cybersecurity

April 23, 2026 By: Jinda Noipho The Conversation

https://theconversation.com/ai-has-crossed-a-threshold-what-claude-mythos-means-for-the-future-of-cybersecurity-281308

he limit of what artificial intelligence can achieve, known as frontier AI, has crossed another threshold. AI can now plan and execute sophisticated cyber operations with minimal guidance at speeds far beyond human capability.

That, at least, is the evidence from an independent test of Claude Mythos Preview, the latest and most advanced model in the Claude family of AI systems, developed by US tech firm Anthropic. Similar to ChatGPT, these can understand and generate human-like text, analyse information, and solve complex problems.

The finance sector is alarmed. It relies on highly interconnected digital systems that are especially attractive targets for sophisticated cyber-attacks. A successful breach could disrupt payments, freeze access to funds, and erode public trust in the banking system.

Major UK and US banks are preparing controlled trials under strict safeguards. They will be granted secure, supervised access to the Mythos Preview model in isolated environments, to evaluate its ability to detect vulnerabilities in their systems while minimising any risk of misuse. It’s a bit like dangerous viruses being examined in high-security laboratories.

The UK’s AI Security Institute, a research organisation within the government’s Department for Science, Innovation and Technology, has already tested Mythos Preview on a demanding benchmark known as The Last Ones. As the name suggests, this series of challenges has been designed as the final hurdle AI systems need to

complete before being deemed able to fully automate complex, real-world cyber-attacks from start to finish.

In the controlled test, Mythos Preview autonomously surfaced thousands of "zero day" vulnerabilities – flaws unknown even to the software's own developers – across every major operating system and popular web browser. Some of these had remained undetected for up to 27 years, even though the software had been carefully checked millions of times.

Under controlled conditions, a skilled human operator would typically need around 20 hours to complete the exercise. In ten independent runs, Mythos achieved full success three times, making this preview version the first AI model to solve the entire attack chain end-to-end.

The results show genuine autonomous chaining of complex sequential actions. Mythos Preview thus represents a major leap in the ability of an AI to act as a truly autonomous agent, planning and executing complex, multi-step tasks over extended periods with minimal human intervention.

But the significance of this technological breakthrough extends well beyond cyber-attacks. The same capability could soon allow AI to autonomously manage software development, scientific research, supply chains or financial operations. Mythos Preview signals a shift from powerful assistant to genuinely autonomous operator, with wide-reaching implications across many industries.

The dual-use dilemma

Rather than releasing it publicly, Anthropic has so far restricted access through its Project Glasswing, an initiative that gives selected technology companies and critical infrastructure providers including Apple, Google, Microsoft, Cisco and Amazon controlled access to the model.

Anthropic's stated idea is to "to secure the world's most critical software" by identifying and fixing security weaknesses in the operating systems, browsers and critical libraries that underpin virtually all modern digital systems, before they can be exploited. Only after that will Mythos see wider deployment as a general-purpose AI system.

Traditional vulnerability management is the process of identifying, assessing and fixing weaknesses in software and systems before attackers can exploit them – a slow, labour-intensive task performed by experts. Mythos could change this process dramatically – in both positive and negative ways.

Its emergence creates a classic dual-use dilemma: the same breakthrough that strengthens cyber defence can also lower the barrier for offensive operations.

On the positive side, it could enable defenders to discover and patch thousands of previously unknown vulnerabilities at unprecedented speed and scale, potentially making critical software far more secure and reducing the window for attacks.

Many current cybercrimes such as ransomware succeed by exploiting known or easily discoverable weaknesses in unpatched systems. These could be significantly reduced if Mythos-class models are widely used for defensive vulnerability discovery.

However, more sophisticated or targeted ransomware attacks – especially those using stolen credentials, social engineering, or already-compromised accounts – are far less likely to be affected, as they often bypass traditional software vulnerabilities altogether.

On the negative side, the same capabilities could dramatically lower the barrier for malicious actors, allowing them to find and chain weaknesses much faster than human teams. This would accelerate sophisticated cyber-attacks if the technology spreads beyond controlled environments.

There is no public evidence that Mythos Preview has reached criminal groups or nation-state adversaries – yet. But the history of cybersecurity technology suggests that well-resourced actors, either state-sponsored or criminal, may develop comparable systems or gain indirect access within the near future.

The future of cybersecurity

In the short term, governments are likely to revise their cybersecurity protocols and incident-response frameworks to incorporate mandatory AI-assisted vulnerability scanning. This would require organisations to continuously scan their systems using AI, rather than relying on occasional human checks.

While this could dramatically improve security by finding flaws faster, it is likely to raise costs significantly and carries the risk of system slowdowns, false alarms, or brief operational disruptions when fixes are applied.

Cyber insurers will almost certainly begin demanding evidence of such defences as a condition of coverage, driving up insurance premiums, while critical-infrastructure operators accelerate deployment of automated monitoring and response systems. This change will impact not only banks and financial institutions, but also critical infrastructure operators in energy, healthcare, telecoms, and transport.

Of course, Mythos is not the final chapter. Future models developed by Anthropic and other leading AI companies are being designed to function as highly autonomous AI agents, capable of independently planning, adapting and executing long, complex sequences of tasks. As well as discovering vulnerabilities, this could mean coordinating large-scale operations or managing sophisticated real-world workflows – all with minimal human guidance.

Moments like this demand both urgency and measured action. Careful governance, international cooperation, and sustained investment in defensive applications will be essential. The genie is out of the bottle – the challenge now is ensuring it serves security rather than chaos.