EUROPEAN POLICY CENTRE

POLICY BRIEF

EUROPE'S POLITICAL ECONOMY PROGRAMME

28 APRIL 2026

Preparedness can't wait: AI pushes cybersecurity into a new era

Paweł Świeboda

INTRODUCTION

Anthropic's reported ability to identify and patch thousands of high-severity software vulnerabilities have shaken not only the technology sector, but also banking and finance. As a result, it has generated unparalleled anxiety among top-level policymakers.

The company's latest advanced model, Claude Mythos, has been shown to be able to identify software bugs that had survived decades of routine review and professional code auditing. In one case, it flagged a 27-year-old flaw during a security audit demonstration. These developments helped push concerns about advanced artificial intelligence (AI) capabilities to the forefront of this year's International Monetary Fund (IMF) and World Bank spring meetings.[1]

A race against time has begun. Political and economic elite are increasingly aware that existing cybersecurity frameworks are ill-equipped for systems with this level of capability. The situation has clear parallels with how drones have transformed conventional warfare, lowering the cost threshold for high-impact attacks.

Within months, access to frontier AI models is likely become much wider. This could lead to a surge of vulnerability reports, making the rapid and decisive patching of software weaknesses a central priority. Companies and public institutions must shift now towards a new cybersecurity posture, with stronger patch management, well-functioning incidence response and effective monitoring as the new normal.

Frontier AI models should be treated as a genuine wake-up call. The question is no longer whether they will transform cybersecurity. It is whether institutions can adapt quickly enough to use them defensively in the face of new risks.

STATE OF PLAY

Enter AI: new speed and scale of finding bugs

Anthropic's new model, Mythos, appears to mark a further step in automated vulnerability discovery, although the broader significance of its capabilities still requires confirmation. Its prededessor, Claude Opus 4.6, had already demonstrated an ability to identify meaningful zero-day vulnerabilities in well-tested codebases. Other notable milestones in this area include Google Project Zero and DeepMind's agent Big Sleep, which identified an exploitable bug in SQLite, the open-source database engine, in late 2024.

That said, Anthropic has not yet published a detailed technical report on Mythos's vulnerability-discovery capabilities or the associated Common Vulnerabilities and Exposures (CVE) identifiers. Nor has the model's performance been independently verified or peer-reviewed. Even so, Mythos appears incrementally more capable than previous models. In particular, it is better able to execute complex workflows in real time, potentially compressing the vulnerability window. While a single zero-day vulnerability can usually be handled through established channels, an AI system that continuously identifies new flaws contributes to a pipeline problem: security teams must rapidly review the output.

Another feature that distinguishes the new AI models is their methodology. Instead of exposing code to a large volume of random inputs, they can reason through code in ways that more closely resemble what human security researchers do. Frontier models can examine older vulnerabilities, analyse how they were patched, and infer where similar weaknesses may persist.

This may improve both the efficiency and the accuracy of vulnerability discovery.

Anthropic presents the model as a defensive tool designed to identify and patch vulnerabilities before they are exploited. This is what many company executives have waited for: numerous organisations would benefit from an AI reviewer that scans through legacy code at speed and helps reduce longstanding backlogs of unpatched vulnerabilities. In general, this can be a positive gain from the point of view of cyber defence.

The same capability, however, also carries clear risks. If a model of this type finds its way into the hands of attackers, it could accelerate the discovery of exploitable entry points in critical infrastructure and other sensitive systems. The risk lies not only in Anthropic's methodology, but also in the wider diffusion of similar capabilities and methods by rogue actors.

The risk lies not only in Anthropic's methodology, but also in the wider diffusion of similar capabilities and methods by rogue actors.

The most advanced AI models are the latest expression of a longer trend: Until recently, many offensive cyber actions had been the domain of experts, whether to write exploit code or deploy attack tools. However, advanced AI is beginning to lower some of those barriers and increase the scale and speed at which such activities can be carried out. As a result, the economics of cybersecurity are changing fast, with AI affecting the cost, immediacy and scale of cyber operations.

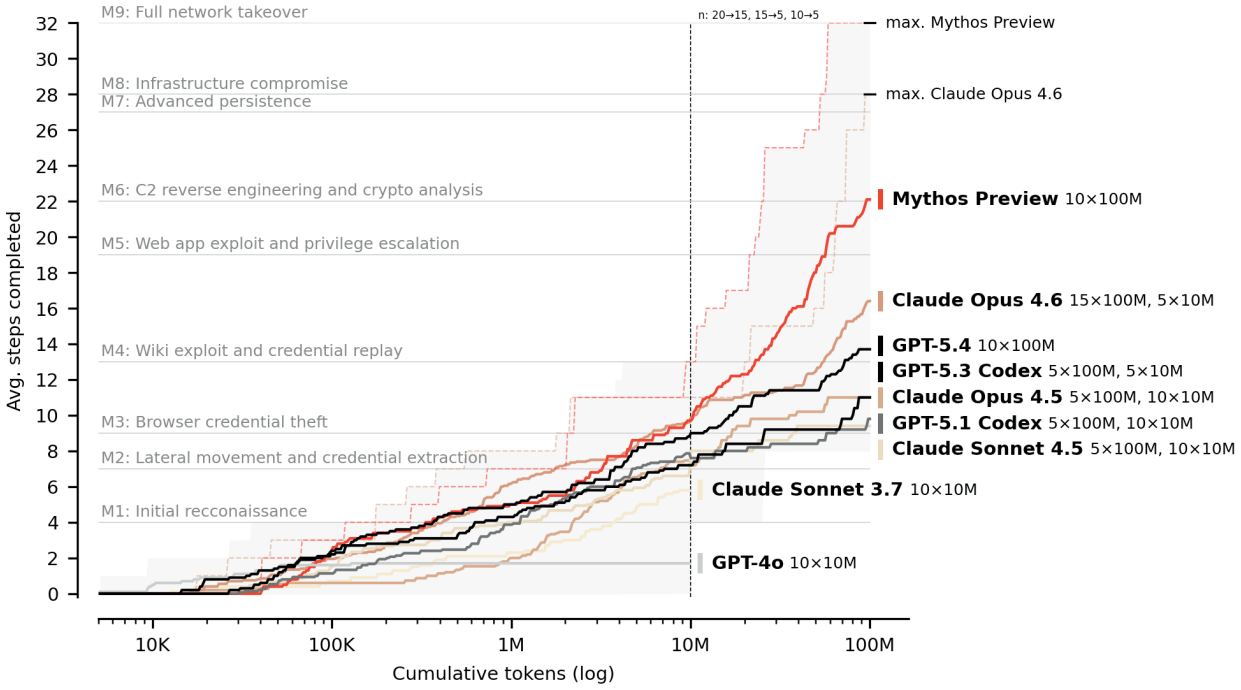
Qualitative leap or incremental improvement?

It is clear that each new generation of frontier models is better at identifying high-severity vulnerabilities, doing so more quickly and without task-specific tooling, custom scaffolding or specialised prompting. At the same time, many questions remain surrounding the extent of Mythos's capabilities. Although concern has reached the top levels of policymaking, Mythos's demonstrated performance may not fully match the company's own claims.

The UK's AI Security Institute's evaluation of Claude Mythos Preview nevertheless suggests a meaningful advance. It concluded that the model represents a "step up over previous frontier models", noting that the landscape was already rapidly improving. As the Institute observed, two years ago, "the best available models could barely complete beginner-level cyber tasks". Today, Mythos Preview is reportedly capable of executing "multi-stage

Figure 1

PERFORMANCE OF AI MODELS MEASURED BY COMPLETED STEPS ON "THE LAST ONES" PER SPENT TOKENS



Source: AI Security Institute, Department of Science, Innovation and Technology of the United Kingdom.

attacks on vulnerable networks and discover and exploit vulnerabilities autonomously".[2]

To its credit, Anthropic had already argued in autumn last year that it was approaching "an inflection point for AI's impact on cybersecurity".[3]

The company's earlier "Threat Intelligence Report"[4] detailed how cybercriminals and other malicious actors are actively attempting to find ways around increasingly sophisticated safety and security measures. It warned that fraudsters have embedded AI throughout all stages of their operations.

Anthropic claims it is on a mission to fix the internet. "Part of tipping the scales towards defenders means doing the work ourselves",[5] the company's experts have said. In this vein, Anthropic is deploying Claude to identify and fix vulnerabilities in open source software. By February 2026, the company had claimed to have identified more than 500 high-severity vulnerabilities, which have begun to be reported. Anthropic works with maintainers to find human-validated bugs and contribute human-reviewed patches.

Offensive versus defensive cybersecurity: A delicate balance

The question of what happens when an AI system finds a hidden weakness in critical software is not new. Goverments and industry have long been aware of the possibility that such capabilities would force difficult choices of who gets access to them, and in what order. However, the issue has not yet received sufficient attention.

Since the onset of the digital revolution, authorities have struggled to respond to the dual-use nature of cyber systems. This tension runs from the "Crypto Wars" of the 1990s over strong encryption to today's frontier AI models, which can help identify zero-day vulnerabilities in widely used software. It is also reflected in debates surrounding the ethics of stockpiling zero-day exploits and their disclosure to vendors for patching. Cybersecurity frameworks have therefore had to balance the public interest in stronger cybersecurity against national security and intelligence demands.

Cybersecurity frameworks have therefore had to balance the public interest in stronger cybersecurity against national security and intelligence demands.

Giving that banning frontier AI models is unlikely, especially as adversaries continue to develop similar capabilities, tight control and monitoring are essential. Primary responsibility rests with the technology companies themselves, whose safeguards are the first line of defence against misuse. Recent threat assessments by the US Cyber Command and allied agencies over the past year have already warned that countries with active cyber-offense programmes are investing heavily in AI-assisted hacking tools. Anthropic's chief executive, Dario Amodei, has likewise argued that open-source models and Chinese developers may be able to replicate Mythos's capabilities within six to 12 months.[6]

Even before the launch of Mythos, Anthropic had introduced a new layer of detection to identify and respond to cyber misuse of its earlier model, Claude. The company says it has introduced new cyber-specific probes to measure activations within the model as it generates a response related to specific harms. It is also updating its cyber enforcement pipelines to keep pace with the new detection architecture. Real-time intervention, including blocking malicious traffic, is now being considered.

The case for making use of these new capabilities for active cyberdefence is strong. Even if the window for exclusive advantage may be closing, securing as much code as possible still has value on its own.

HOW FAST IS FAST ENOUGH?

Anthropic's Mythos offers the possibility of rapidly shrinking patch cycles and hence the window for attackers to exploit existing weaknesses.

Today's cybersecurity frameworks are fast but do not move at the speed of light. The Vulnerabilities Equities Process used by the US government to decide whether to disclose or retain zero-day vulnerabilities in information systems involves the Equities Review Board, which meets monthly. The system was designed for individual vulnerabilities, rather than for a system which may act across a range of bugs in many codebases at the same time. This means that existing frameworks must be rapidly adjusted to absorb AI-scale discovery.

Urgency of EU-level action

The EU's AI Office has held its cards close to its chest on the risks inherent to advanced AI models. It has yet to carry out a risk assessment akin to that of the UK AI Security Office, prompting a coalition of eight AI safety groups to call on the Commission to prop up its expertise on the issue.[7] They draw attention to the fact that the less than two-years-old EU AI Office has around 140 staffers, of which just 36 are in the safety unit. Although Commission spokesperson Thomas Regnier argued that the "AI Office has built state-of-the-art model evaluation capacity",[8] he has also confirmed that the institution was not given full early access to the Mythos model, unlike 40 other organisations.

The EU's revised Cybersecurity Act acknowledges that "emerging technologies like the artificial intelligence (AI) and quantum computing are reshaping the tools of defence and the tactics of adversaries".[9] However, the proposed solutions do not prioritise addressing the rapid advances of frontier models. Neither does the Act mention advanced AI systems among the main problems that the revision of the CSA aims to tackle. It focuses instead on the misalignment between the Union's cybersecurity policy framework and stakeholder needs, stalled implementation of the European cybersecurity certification framework, the complexity of cybersecurity-related policies, and increasing Information and Communication Technologies (ICT) supply chains security risks.

RECOMMENDATIONS

While the anxiety surrounding the new capabilities of frontier AI models is understandable, there remains scope for meaningful policy action. Assuming that the new capabilities will only improve and diffuse over time, the focus should be squarely on managing their impact, not trying to put the genie back in the bottle.

Four priorities stand out:

1. Pre-deployment assessment of frontier AI models

The release of new frontier AI models should be subject to clearly defined criteria and independent assessment before commercial deployment. This should include rigorous evaluation of their capacity to discover and potentially exploit vulnerabilities.

Such a process should establish standards and procedures for the safe and effective defensive use of frontier AI, including how vulnerability discovery and rapid response can be operationalised. The central challenge is not only detection, but whether organisations can act upon that information safely, immediately and effectively.

2. Controlled release of frontier models

The launch of new high-capability models should be accompanied by structured cooperation with trusted partners, which would include joint testing of the capabilities. Anthropic's Project Glasswing, which brings together more than 40 organisations including Amazon, Apple and Microsoft to identify and patch vulnerabilities, offers one possible model. However, the formula needs to be more comprehensive and balanced, not creating privileged access and hence competitive advantage to some companies. It should also include public institutions responsible for AI safety such as the EU's AI Office. More broadly, the expert community needs to move quickly in adapting responsible-disclosure norms to the age of AI-assisted auditing without releasing methodological details that could benefit adversaries.

Although one immediate option could be to restrict access to highly capable models by introducing a licensing regime overseen by the relevant public authority, the downsides of this approach prevail over possible advantages. Such a framework could classify high-risk systems, limit access to approved users and monitor compliance. This solution would mirror the existing practice of retaining a stock of classified vulnerability information within intelligence agencies. The drawback, however, would be that it could reproduce an already suboptimal siloed situation, in which many civilian infrastacture operators are excluded from the defensive capabilities of the new systems.

3. New industry cybersecurity standards and workflows

The risk of frontier AI models being used by malicious actors is especially acute for organisations with weak cybersecurity preparedness. Standards and workflows therefore need to be adjusted to reflect the possibility that large numbers of LLM-discovered bugs may need to be patched rapidly.

This requires businesses and other actors to raise their security baselines. The European Commission should develop stronger collaboration with industry to reach that defensive depth. It also means that existing disclosure norms must reflect this new sense of urgency.

Many security teams have already invested in automated vulnerability discovery to identify risks at scale. What is changing now is the speed and volume, meaning they must be ready for realtime triage and validation. The main danger is that organisations will not be able to respond fast enough. If dealing with thousands of vulnerabilities at once becomes possible, patch velocity may emerge as a new, competitive capacity organisations will strive to achieve.

CONCLUSION

Extreme solutions such as banning the rollout of AI models that are increasingly capable of finding high-severity vulnerabilities at scale would not only be operationally difficult, but also counterproductive. The EU should seek instead to harness frontier models' cyber-defence potential while doing the utmost to limit access by adversaries.

Advanced AI has now pushed cybersecurity to the centre of the economic security debate. Policymakers face a strategic decision between trying to constrain or control the rollout of powerful AI models, or rapidly deploying them for defence.

In either scenario, the priority is to understand the threat landscape clearly and stay ahead of adversaries. The future of AI-powered cybersecurity is unsettled, but the stakes are already clear.

Paweł Świeboda is a Senior Visiting Fellow and Co-Director of the Brussels Economic Security Forum.

NOTES

NOTES

[1] "Latest AI models could threaten world banking system, financial officials warn", *Financial Times*.

[2] AI Security Institute (AISI), "Our evaluation of Claude Mythos Preview's cyber capabilities", 13 April 2026.

[3] "Building AI for cyber defenders", red anthropic, 29 September 2025.

[4] Moix, Alex; Lebedev, Ken and Klein, Jacob (2025), Threat Intelligence Report: August 2025, Anthropic.

[5] Carlini, Nicholas et al, "Evaluating and mitigating the growing risk of LLM-discovered 0-days", red anthropic, 5 February 2026.

[6] "Anthropic chief Dario Amodei: 'I don't want AI turned on our own people'", *Financial Times*.

[7] Pour Demain NGO, Open joint letter on AI Office resourcing, 15 April 2026.

[8] Haeck, Pieter and Clark, Sam, "Anthropic's hacking tech exposes EU AI Office weaknesses", POLITICO, 16 April 2026.

[9] European Commission, Proposal for a Regulation for the EU Cybersecurity Act, ec.europa.eu, 20 January 2026.

With the strategic support of King Baudouin Foundation
Working together for a better society