National Cyber and Information Security Agency

NÚKIB

BRNO • JULY 7 2026
[THREAT ANALYSIS]

THE NEW AI FRONTIER MODELS SIGNIFICANTLY ENHANCE THE ABILITY TO AUTONOMOUSLY DETECT AND EXPLOIT VULNERABILITIES: IT IS ONLY A MATTER OF TIME BEFORE MALICIOUS ACTORS EXPLOIT THEM

SUMMARY

- **In April 2026, Anthropic unveiled Mythos, an artificial intelligence (AI) model that, thanks to advanced automation, is capable of searching for, verifying, and exploiting even previously unknown vulnerabilities on a large scale—and doing so significantly more effectively than older large language models. This capability poses a real and serious risk to cybersecurity.**
- **This represents a shift across the entire AI industry, in which other companies will soon acquire similar capabilities for identifying and exploiting vulnerabilities. It is almost certain (90–100%) that these capabilities will also fall into the hands of state and non-state adversaries, if they do not already possess them.**
- **These advanced Frontier AI models, when applied to cybersecurity, are very likely (75–85%) to significantly reduce the time between the discovery of a vulnerability and its exploitation, as they automate the search for, verification of, and chaining of vulnerabilities on a large scale. However, the pace of patching will continue to be limited by human and technical capacities.**
- **In the short term, new AI models may increase the number of cyber incidents inCzechia as well, both through direct attacks on Czech entities and indirectly through global compromises of major service providers via supply chains.**
- **RECOMMENDATION: These Frontier AI models will significantly speed up and expand the search for existing vulnerabilities. Effective defense will depend on efficient vulnerability reporting and handling systems, rapid prioritization of the highest-risk vulnerabilities, and, later on, the use of these tools to test your own infrastructure.**

NOTICE: The information and conclusions contained in this analysis are based on publicly available information and on information obtained through National Cyber and Information Security Agency (NÚKIB) activities at the time of publication. This is an analysis of cybersecurity from NÚKIB's perspective, based on the information available to it.

A new generation of large language models (LLMs), known as cybersecurity-focused frontier models, brings transformative capabilities to the field of cybersecurity, particularly in the context of detecting vulnerabilities and their exploitation. **These new models can automatically scan code at a scale that is difficult for human security teams to manage, identify errors, and simultaneously generate hypotheses about their potential exploitability.**[1] Their capabilities also include automated validation of identified weaknesses and the ability to chain together a series of lower-severity vulnerabilities so that, cumulatively, they become critical.

The first model of this new generation, Mythos from Anthropic, reportedly successfully exploited newly discovered vulnerabilities in approximately 72% of test cases and was able to autonomously discover previously unknown vulnerabilities in operating systems and web browsers.[2] Anthropic stated that the system can identify even very hard-to-detect vulnerabilities, including those that have persisted in software for decades. The company has made the Mythos model available as part of the Glasswing project only to a limited group of U.S. partners (such as Microsoft, Google, Apple, and Amazon).[3]

Anthropic subsequently made the model available to the general public on June 9 as Claude Fable 5, which

was designed to include robust safeguards against misuse.[4] **However, it subsequently withdrew the app again after U.S. authorities expressed concerns about the possibility of circumventing the model's security restrictions.**[5] According to available information, Fable 5 had not been relaunched as of June 16.

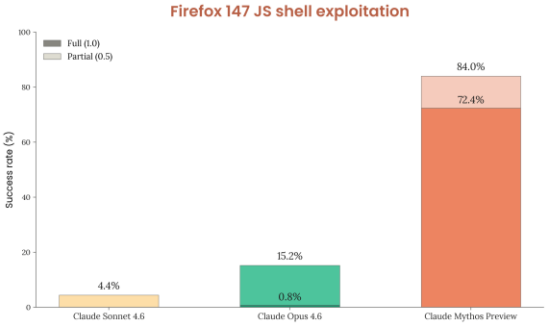
Most leaders in the field of AI are, of course, developing similar models. At present, however, it is unclear how and when they will be released. **It is almost certain (90–100%) that, within a year at the latest, additional commercial models with these capabilities will be widely introduced in the field of cybersecurity.**

AI WILL SHORTEN THE TIME IT TAKES TO EXPLOIT VULNERABILITIES, BUT THE TIME NEEDED TO FIX THEM WILL REMAIN THE SAME FOR NOW

Frontier AI models focused on cybersecurity are very likely (75–85%) to benefit cybersecurity in the long term, **but in the short term, they also pose a significant cybersecurity risk.**

For the time being, at least, vulnerability fixes will continue to be limited by human, technological, and infrastructural capacities. **Until this imbalance is corrected (for example, through deeper integration of AI tools into the process of identifying and fixing vulnerabilities), these models will provide an advantage to both state and non-state attackers.** They will be able to automate a much larger portion of their activities, which **will lead, among other things, to an increase in their operational pace and make these operations more difficult to detect**.[6]

Obr. 1: The success rate of the Mythos model in detecting vulnerabilities in the Firefox browser, compared to previous models from Anthropic (higher resolution)



Source: vellum.ai

The ability of these Frontier AI models to effectively exploit vulnerabilities then further deepens the existing, long-standing cybersecurity problem posed by legacy technologies. [7] Companies often struggle with outdated and inadequately secured legacy software or hardware that, for various reasons, they cannot update or replace (e.g., because the vendor is no longer in business). These systems often suffer from accumulated technical and operational shortcomings that make them difficult to repair, and any changes to them pose a high risk of disrupting their operation.[8] **Operational technologies (OT) pose a major challenge in this regard.**[9] **Furthermore, the vulnerabilities found in them usually take longer to fix.**

IN THE SHORT TERM, WE CAN EXPECT MORE COMMERCIAL AND OPEN-SOURCE MODELS

Although attention is now focused primarily on the Mythos model and Anthropic's measures to prevent its misuse, **it is important to note that this represents a systemic shift in the capabilities of AI models in general**. With previous milestones (such as video generation), competing companies have always achieved capabilities comparable to those of the leader at the time within a short period.[10] For example, both Google and OpenAI are currently developing models with similar capabilities (Big Sleep and GPT-5.5-Cyber).[11] **It is almost certain (90–100%) that Chinese AI companies are also working on this type of Frontier AI model with similar capabilities, if they do not already have them.**[12]

Although AI companies are currently taking a limited approach to these models, focusing on identifying vulnerabilities to benefit defenders (e.g., accelerating the penetration testing process), **it is almost certain (90–100%) that, driven by competition, some of these companies will make their models widely available commercially in the short term**. Furthermore, similar results can be achieved even with less advanced, modified models.[13]

With LLM, it is also possible to take advantage of the principle of distillation. This is a process in which a smaller model acquires the knowledge and behavior of a larger, more powerful model by being trained on its outputs.[14] The United States has previously accused the People's Republic of China (PRC) of using this process to close the gap with U.S. companies in the field of AI.[15] **It is therefore highly likely (75–85%) that**

these capabilities will become widely available in the short term (primarily through the misuse of publicly available commercial services). At the same time, it is likely (55–70%) that state-sponsored actors from countries with advanced AI sector may already possess these capabilities today.

IMPLICATIONS FOR THE CZECH REPUBLIC

In the short term, Czechia may also see a significant increase in incidents related to the capabilities of these types of Frontier AI models. It is highly likely (75–85%) that this massive increase in the number of discovered vulnerabilities will be difficult to monitor. This will have a direct impact on the increased workload of security teams responsible for processing incident reports.

These may be direct attacks on Czech entities; however, the most likely scenario involves the repercussions of breaches at large companies providing global services, which will also have a direct impact on the Czech Republic through the supply chain. For example, attacks could target the shared digital infrastructure of banks, payment systems, and cloud and software service providers.[16] We can also expect an increase in the number of updates as the current code is reviewed for the first time using new AI models and new vulnerabilities are discovered in greater numbers. This will impose increased demands on the application of updates.

RECOMMENDATION

Mythos and other Frontier AI models do not introduce fundamentally new methods for identifying and exploiting vulnerabilities, but they significantly increase the speed and scalability of these processes.

It is highly likely (75–85%) that systems and processes for addressing vulnerabilities will soon be overwhelmed by the volume of newly discovered vulnerabilities, so effective defense will depend on strict adherence to cybersecurity policies, proper prioritization, and rapid assessment of which vulnerabilities pose a real risk to a specific organization.

At the same time, we must anticipate that attackers will likely need less time to exploit disclosed vulnerabilities; therefore, we will need to accelerate the deployment of patches, while always taking into account system stability and operational impacts.

Once these models become publicly available, it will also be advisable to integrate them into your security processes and use them to test your own infrastructure and identify vulnerabilities in your own environment.

Software developers should therefore, where economically and operationally possible, use AI-based reviews to identify vulnerabilities before their products are released to the market.

SOURCES

[1] The Register. 2026. Anthropic Mythos model can find and exploit 0-days. https://www.theregister.com/security/2026/04/08/anthropic-mythos-model-can-find-and-exploit-0-days/5224393, Anthropic. 2026. Project Glasswing: Securing critical software for the AI era. https://www.anthropic.com/glasswing.

[2] SOCFortress. 2026. The Mythos Singularity: Why Cyber Defense Just Lost the Luxury of Time. https://socfortress.medium.com/the-mythos-singularity-why-cyber-defense-just-lost-the-luxury-of-time-3c2ce65dbb7e.

[3] Anthropic. 2026. Project Glasswing: Securing critical software for the AI era. https://www.anthropic.com/glasswing.

[4] Anthropic, "Claude Fable 5 and Claude Mythos 5," *Anthropic News*, June 9, 2026, https://www.anthropic.com/news/claude-fable-5-mythos-5

[5] Anthropic, "Statement on the US Government Directive to Suspend Access to Fable 5 and Mythos 5," *Anthropic Announcements*, June 12, 2026, https://www.anthropic.com/news/fable-mythos-access.

[6] SOCFortress. 2026. The Mythos Singularity: Why Cyber Defense Just Lost the Luxury of Time. https://socfortress.medium.com/the-mythos-singularity-why-cyber-defense-just-lost-the-luxury-of-time-3c2ce65dbb7e.

[7] ISA Global Cybersecurity Alliance, "Addressing Cybersecurity Risks in Legacy OT Systems: A Practical Guide," *Automation.com*, January 2024, https://www.automation.com/article/cybersecurity-risks-legacy-ot-systems

[8] Clothier, Mat. 2025. "Why Are Companies Not Tackling Their Windows Technical Debt?" *TechRadar Pro*, December 6, 2025. Accessed August 3, 2026. https://www.techradar.com/pro/why-are-companies-not-tackling-their-windows-technical-debt

[9] Asset Guardian. 2025. "OT Patch Management: How to Secure Systems You Can't Patch." *Asset Guardian Insights*. Accessed August 3, 2026. https://www.assetguardian.com/insights/insights-ot-patch-management-when-you-cant-patch-legacy-systems

[10] Field, Hayden. 2024. OpenAI releases Sora, its buzzy AI video-generation tool. https://www.cnbc.com/2024/12/09/openai-releases-sora-its-buzzy-ai-video-generation-tool.html, Neural Frames. 2025. Seedance 1.0: ByteDance's Lightning-Fast AI Video Engine and Why the Music Video World Should Pay Attention. https://www.neuralframes.com/post/seedance-1-0-bytedances-lightning-fast-ai-video-engine-and-why-the-music-video-world-should-pay-attention, MindStudio. 2025. What Is Google Veo 2? AI Video Generation Explained. https://www.mindstudio.ai/blog/what-is-google-veo-2-video-generation

[11] Google Project Zero. 2024. From Naptime to Big Sleep: Using Large Language Models To Catch Vulnerabilities In Real-World Code. https://projectzero.google/2024/10/from-naptime-to-big-sleep.html, OpenAI. 2026. Scaling Trusted Access for Cyber with GPT-5.5 and GPT-5.5-Cyber. https://openai.com/cs-CZ/index/gpt-5-5-with-trusted-access-for-cyber/

[12] Cerulus, Laurens. 2026. China is going dark to develop its own Mythos, German cyber chief fears. https://www.politico.eu/article/china-is-going-dark-to-develop-its-own-mythos-german-cyber-chief-fears/

[13] Kim, Taesoo. 2026. "Defense at AI Speed: Microsoft's New Multi-Model Agentic Security System Tops Leading Industry Benchmark." *Microsoft Security Blog*, May 12, 2026. Microsoft. Accessed August 3, 2026. https://www.microsoft.com/en-us/security/blog/2026/05/12/defense-at-ai-speed-microsofts-new-multi-model-agentic-security-system-tops-leading-industry-benchmark/, AISLE. 2026. AI Cybersecurity After Mythos: The Jagged Frontier. https://aisle.com/blog/ai-cybersecurity-after-mythos-the-jagged-frontier.

[14] MindStudio. 2026. AI Model Distillation Attacks Explained: What They Are and Why They Matter. https://www.mindstudio.ai/blog/ai-model-distillation-attacks-explained

[15] BBC News. 2026. White House memo claims mass AI theft by Chinese firms. https://www.bbc.com/news/articles/cpqxgxx9nrqo

[16] International Monetary Fund. 2026. Financial Stability Risks Mount as Artificial Intelligence Fuels Cyberattacks. https://www.imf.org/en/blogs/articles/2026/05/07/financial-stability-risks-mount-as-artificial-intelligence-fuels-cyberattacksod

TERMS OF USE OF INFORMATION

The use of the information provided shall be in accordance with the Traffic Light Protocol methodology. The information is marked with a flag that specifies the conditions of use of the information. The following flags are set, indicating the nature of the information and the conditions of use:

Colour	Conditions of use
Red **TLP:RED**	The information may not be provided to any person other than the person to whom the information was addressed unless other persons to whom such information may be provided are specifically identified. Where the recipient considers it important to disclose the information to other bodies, this may be done only with the consent of the originator of the information.
Orange **TLP:AMBER+STRICT**	The information may only be shared within the recipient's organisation, and only to persons who meet the need-to-know and whose information is relevant to resolving the problem or threat identified in the information.
Orange **TLP:AMBER**	Information may be shared within the recipient organisation and to its partners, and only to persons who meet the need-to-know and whose information is relevant to resolving the problem or threat identified in the information.
Green **TLP:GREEN**	Information may be shared within the beneficiary's organisation and, where appropriate, with other partners of the beneficiary, but not through publicly available channels; the beneficiary must ensure the confidentiality of the communication when forwarding it.
TLP:CLEAR	The information may be further provided and disseminated without restriction. Any restrictions based on the intellectual property rights of the originator and/or recipient or third parties are not affected by this provision.

EXPRESSION OF NÚKIB PROBABILITIES

Expression	Probability
Almost sure	*90-100 %*
Very likely	*75-85 %*
Probable	*55-70 %*
Cannot be ruled out/Real possibility	*40-50 %*
Improbable	*20-35 %*
Very unlikely	*0-15 %*