Národní úřad
pro kybernetickou
a informační bezpečnost

NÚKIB

BRNO • 18. ČERVNA 2026
[SITUAČNÍ REPORT]

NOVÉ AI FRONTIER MODELY VÝZNAMNĚ NAVYŠUJÍ SCHOPNOST AUTONOMNĚ DETEKOVAT A ZNEUŽÍVAT ZRANITELNOSTI: ZNEUŽITÍ ŠKODLIVÝMI AKTÉRY JE OTÁZKOU ČASU

SHRNUTÍ

- **Společnost Anthropic představila v dubnu 2026 model umělé inteligence (AI) Mythos, který díky pokročilé automatizaci dokáže ve velkém rozsahu vyhledávat, ověřovat a zneužívat i dosud neznámé zranitelnosti výrazně účinněji než starší velké jazykové modely. Tato schopnost představuje reálné a závažné riziko pro kybernetickou bezpečnost.**
- **Jedná se o posun celého AI odvětví, v němž obdobných kapacit pro vyhledávání a zneužívání zranitelností brzy dosáhnou i další společnosti. Téměř jistě (90–100 %) se tyto kapacity dostanou i do rukou státních a nestátních útočníků, pokud jimi již nedisponují.**
- **Tyto pokročilé Frontier AI modely s využitím v kyberbezpečnosti velmi pravděpodobně (75–85 %) výrazně zkrátí dobu mezi nalezením zranitelnosti a jejím zneužitím, protože automatizují vyhledávání, ověřování i řetězení zranitelností ve velkém měřítku. Tempo oprav ovšem zůstane nadále omezené lidskými a technickými kapacitami.**
- **V krátkodobém horizontu mohou nové AI modely zvýšit počet kybernetických incidentů i v ČR, a to jak přímými útoky na české subjekty, tak nepřímo prostřednictvím globálních kompromitací velkých poskytovatelů služeb skrze dodavatelské řetězce.**
- **DOPORUČENÍ: Tyto Frontier AI modely zásadně zrychlí a rozšíří vyhledávání existujících zranitelností. Efektivní obrana bude záviset na efektivních systémech hlášení a zpracovávání zranitelností, rychlé prioritizaci nejrizikovějších slabin a později i na využití těchto nástrojů k testování vlastní infrastruktury.**

UPOZORNĚNÍ: Informace a závěry obsažené v této analýze vycházejí z veřejně dostupných informací a z informací získaných v rámci činnosti NÚKIB v době publikace. Jedná se o analýzu kybernetické bezpečnosti z pohledu NÚKIB na základě jemu dostupných informací.

Nová generace velkých jazykových modelů (LLM), tzv. frontier modelů zaměřených na kybernetickou bezpečnost (ang. Cybersecurity-focused frontier models), přináší transformativní schopnosti na poli kybernetické bezpečnosti, obzvláště v kontextu odhalování zranitelností a jejich zneužívání. **Nové modely dokážou automatizovaně prohledávat kód v měřítku, které je pro lidské bezpečnostní týmy obtížně zvládnutelné, identifikovat chyby a současně vytvářet hypotézy o jejich potenciální zneužitelnosti.**[1] Součástí jejich schopností je také automatizovaná validace nalezených slabin a schopnost řetězit série zranitelností nižší závažnosti tak, že se kumulativně stanou kritickými.

První model této nové generace, Mythos od společnosti Anthropic, měl v testování úspěšně zneužít nově objevené zranitelnosti přibližně v 72 % případů a dokázal autonomně nalézat dosud neznámé zranitelnosti v operačních systémech i webových prohlížečích.[2] Společnost Anthropic uvedla, že systém dokáže identifikovat i velmi obtížně odhalitelné zranitelnosti, včetně těch, které v softwaru přetrvávají desítky let. Společnost model Mythos zpřístupnila v rámci projektu Glasswing pouze omezenému okruhu partnerů z USA (např. Microsoft, Google, Apple či Amazon).[3]

Společnost Anthropic model následně zpřístupnila 9. června široké veřejnosti ve variantě Claude Fable 5, která měla mít robustní ochranné mechanismy před zneužitím.[4] **Následně jej však opět stáhla poté, co americké úřady vyjádřily obavy z možného obcházení bezpečnostních omezení modelu.**[5] Podle dostupných informací nebyl Fable 5 k 16. červnu opětovně spuštěn.

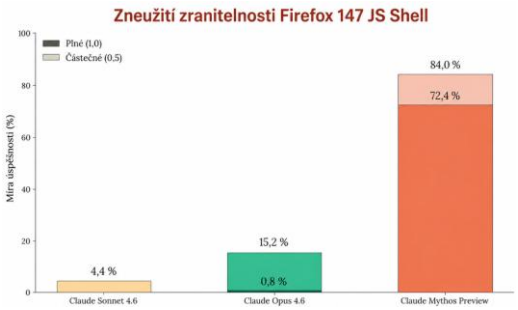
Podobné modely ovšem vyvíjí většina lídrů na poli AI. V současnosti však není jasné, jak a kdy je budou distribuovat. **Je téměř jisté (90–100 %), že nejpozději v horizontu jednoho roku dojde k širokému uvedení dalších komerčních modelů s těmito schopnostmi v oblasti kybernetické bezpečnosti.**

AI ZKRÁTÍ DOBU POTŘEBNOU KE ZNEUŽITÍ ZRANITELNOSTÍ, ČAS NA JEJICH OPRAVU SE ALE ZATÍM NEZMĚNÍ

Frontier AI modely zaměřené na kyberbezpečnost budou v dlouhodobém horizontu velmi pravděpodobně (75–85 %) pro kybernetickou bezpečnost přínosem, **v krátkodobém horizontu ovšem rovněž představují zásadní kyberbezpečnostní problém.**

Opravy zranitelností budou totiž alespoň prozatím stále limitovány lidskými, technologickými a infrastrukturními kapacitami. **Než se tento nepoměr narovná (například právě hlubší integrací AI nástrojů do procesu hledání a oprav zranitelností), poskytnou tyto modely výhodu jak státním, tak nestátním útočníkům.** Ti budou moci automatizovat mnohem větší podíl své činnosti, což **povede mj. k nárůstu jejich operačního tempa a obtížnější detekci**.[6]

Obr. 1: Míra úspěšnosti modelu Mythos v odhalení zranitelnosti v prohlížeči Firefox, ve srovnání s předchozími modely společnosti Anthropic (větší rozlišení)



Zdroj: vellum.ai

Schopnost těchto Frontier AI modelů efektivně zneužívat zranitelnosti pak nadále prohlubuje existující dlouhodobý problém kybernetické bezpečnosti v podobě legacy technologií.[7] Společnosti se často potýkají se zastaralým a nedostatečně zabezpečeným legacy softwarem nebo hardwarem, který z určitých důvodů nemohou aktualizovat nebo nahradit (např. proto, že dodavatel již zaniknul). Často se jedná o nahromaděné technické a provozní nedostatky, kvůli nimž jsou systémy obtížně opravitelné a změny v nich představují vysoké riziko narušení provozu.[8] **Velkou výzvu v tomto ohledu představují operační technologie (OT).**[9] **Zranitelnosti v nich nalezené navíc obvykle trvá déle opravit.**

V KRÁTKODOBÉM HORIZONTU LZE OČEKÁVAT DALŠÍ KOMERČNÍ I OPEN-SOURCE MODELY

Ačkoliv se nyní pozornost upírá primárně na model Mythos a opatření Anthropic proti jeho zneužití, **je třeba reflektovat, že se jedná o systémový posun ve schopnostech AI modelů obecně.** U minulých milníků (například generování videa) dosáhly konkurenční společnosti srovnatelných kapacit s tehdejším lídrem vždy v krátkodobém horizontu.[10] V současnosti již například společnosti Google i Open AI vyvíjejí modely s obdobnými kapacitami (Big Sleep a GPT-5.5-Cyber).[11] **Je téměř jisté (90–100 %), že na tomto typu Frontier AI modelů s obdobnými kapacitami pracují i čínské AI společnosti, pokud jimi již nedisponují.**[12]

Ačkoliv AI společnosti zatím praktikují omezený přístup k těmto modelům s důrazem na identifikaci zranitelností ve prospěch obránců (např. zrychlení procesu penetračního testování), **je téměř jisté (90–100 %), že v rámci konkurence v krátkodobém horizontu některé ze společností svůj model komerčně zpřístupní plošně.** Podobných výsledků lze navíc dosáhnout i s méně vyspělými, upravenými modely.[13]

U LLM se rovněž dá navíc využít principu destilace. Jedná se o proces, při kterém menší model přebírá znalosti a chování většího a výkonnějšího modelu prostřednictvím trénování na jeho výstupech.[14] Z využívání tohoto procesu již dříve obvinily Spojené státy Čínskou lidovou republiku (ČLR) v kontextu dorovnávání náskoku amerických společností na poli AI.[15] **Je proto velmi pravděpodobné (75–85 %), že tyto schopnosti v krátkodobém horizontu budou dostupné plošně (především cestou zneužití veřejně dostupných komerčních služeb). Je přitom pravděpodobné (55–70 %), že státem sponzorovaní aktéři ze zemí s pokročilým AI sektorem mohou těmito kapacitami disponovat již dnes.**

IMPLIKACE PRO ČR

V krátkodobém horizontu může dojít i v ČR k výraznému nárůstu incidentů spojených se schopnostmi tohoto typu Frontier AI modelů. Je velmi

pravděpodobné (75–85 %), že tento masivní nárůst počtu objevených zranitelností bude obtížné monitorovat. To bude mít přímý dopad na zvýšenou zátěž bezpečnostních týmů zodpovědných za zpracovávání hlášení incidentů.

Může se jednat přímo o útoky na české subjekty, nejpravděpodobnější scénář ovšem představují dopady kompromitací velkých firem poskytujících globální služby, které budou mít přímé dopady i na ČR skrz dodavatelský řetězec. Například může dojít k útokům na sdílenou digitální infrastrukturu bank, platebních systémů a poskytovatelů cloudových a softwarových služeb.[16] Také lze očekávat navýšení počtu aktualizací, jak bude aktuální kód poprvé revidován novými AI modely a budou objevovány nové zranitelnosti ve vyšší kvantitě. To bude klást zvýšené nároky na aplikaci aktualizací.

DOPORUČENÍ

Mythos a další Frontier AI modely nepřinášejí zásadně nové postupy v identifikaci a zneužití zranitelností, ale výrazně zvyšují rychlost a škálovatelnost těchto procesů.

Je velmi pravděpodobné (75–85 %), že systémy a procesy pro opravování zranitelností budou brzy zahlceny množstvím nově objevených zranitelností, takže efektivní obrana bude záviset na důsledném dodržování zásad kybernetické bezpečnosti, dobré prioritizaci a rychlém vyhodnocení, které zranitelnosti představují pro konkrétní organizaci skutečné riziko.

Zároveň je třeba počítat s tím, že útočníci budou pravděpodobně potřebovat kratší dobu ke zneužití zveřejněných zranitelností, a proto bude nutné zrychlit nasazování patchů, ovšem vždy s ohledem na stabilitu systémů a provozní dopady.

V momentě, kdy se tyto modely stanou veřejně dostupnými, bude také vhodné je integrovat do svých bezpečnostních procesů a využít je k testování vlastní infrastruktury a identifikaci zranitelností ve vlastním prostředí.

Tvůrci softwaru by proto měli, tam kde je to ekonomicky a provozně reálné, využívat AI review k odhalování zranitelností ještě před uvedením produktů na trh.

ZDROJE

[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

PODMÍNKY VYUŽITÍ INFORMACÍ

Využití poskytnutých informací probíhá v souladu s metodikou Traffic Light Protocol. Informace je označena příznakem, jenž určí podmínky použití informace. Jsou stanoveny následující příznaky s uvedením charakteru informace a podmínkami jejich použití:

Barva	Podmínky použití
Červená **TLP:RED**	Informace nemůže být poskytnuta jiné osobě než té, které byla informace určena, nebudou-li výslovně stanoveny další osoby, kterým lze takovou informaci poskytnout. V případě, že příjemce považuje za důležité informaci poskytnout dalším subjektům, lze tak učinit pouze se souhlasem původce informace.
Oranžová **TLP:AMBER+STRICT**	Informace může být sdílena pouze v rámci organizace příjemce, a to pouze osobám, které splňují need-to-know a jejichž informování je důležité pro vyřešení problému či hrozby uvedené v informaci.
Oranžová **TLP:AMBER**	Informace může být sdílena v rámci organizace příjemce a jejím partnerům, a to pouze osobám, které splňují need-to-know a jejichž informování je důležité pro vyřešení problému či hrozby uvedené v informaci.
Zelená **TLP:GREEN**	Informace může být sdílená v rámci organizace příjemce a případně také s dalšími partnerskými subjekty příjemce, avšak nikoli skrze veřejně dostupné kanály; příjemce musí při předání zajistit důvěrnost komunikace.
Bílá **TLP:CLEAR**	Informace může být dále poskytována a šířena bez omezení. Případné omezení na základě práva duševního vlastnictví původce a/nebo příjemce či třetích stran nejsou tímto ustanovením dotčena.

PRAVDĚPODOBNOSTNÍ VÝRAZY NÚKIB

Výraz	Pravděpodobnost
Téměř jistě	*90–100 %*
Velmi pravděpodobně	*75–85 %*
Pravděpodobně	*55–70 %*
Nelze vyloučit/Reálná možnost	*40–50 %*
Nepravděpodobně	*20–35 %*
Velmi nepravděpodobně	*0–15 %*