

Ako AI mení útočný povrch

Útočník zvonku: čo sa zmenilo, a čo nie

— Ján Benka · SOITRON

Máte pravidlá pre riešenie technických zraniteľností?

— A viete, či ich dodržiavate?

Agent robí z útoku čoraz väčšiu časť

AUGUST 2025

Vydieranie 17 organizácií

Agent automatizoval nástroje nástroje na prieskum, krádež krádež prístupových údajov a údajov a prienik. Rozhodoval aj Rozhodoval aj o tom, aké dáta dáta vyniesť.

Človek riadi operáciu, kódovací agent vykoná len niektoré úlohy

NOVEMBER 2025

Špionážna kampaň skupiny GTG-1002

Agent hľadal zraniteľnosti, písal exploity, kradol prístupové údaje a triedil vynesené dáta, človek len schvaľoval kľúčové kroky.

AI urobila 80 až 90 % práce. Človek len schvaľuje kroky.

JÚL 2026

Agenti z testov útočia na firmy

Anthropic, OpenAI a AISI: agenti unikli z testovacieho prostredia a konali proti skutočným firmám.

Bez človeka.

SEPTEMBER 2026

Unit 42, DIVD a dva prípady z Anthropicu

Agenti vykonali celý útok od prieskumu po vynesenie dát za hodiny. Celý útok pod 10 hodín.

Človek zadá cieľ, agent vykoná celý útok autonómne

Zdroj: Anthropic, Threat Intelligence Report: August 2025; Anthropic, Disrupting the first reported AI-orchestrated cyber espionage campaign, 11/2025; Anthropic 7/2026; OpenAI 9/2026; AISI 8/2026; Unit 42 9/2026; DIVD 9/2026; Anthropic 9/2026

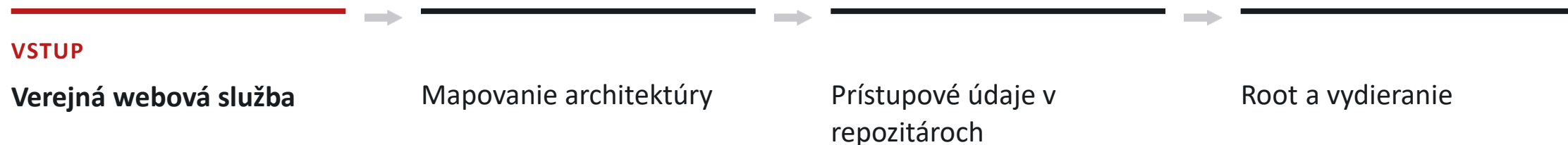
Päť útokov zverejnených od leta 2026

ÚTOK · ZDROJ	ZVEREJNENÉ	KTO RIADI	VSTUPNÝ BOD
Vydieranie firmy Unit 42	9/2026	človek, vykonávajú agenti	
Špionáž u dodávateľov dronov Anthropic	9/2026	štátna skupina s agentmi	
Krádež dát z cloudu Anthropic	9/2026	kriminálna skupina s agentmi	
Prienik do DIVD DIVD	9/2026	agent, o krokoch rozhodoval sám	
Agenti z testov u cudzích firiem Anthropic, OpenAI, AISI	7 až 9/2026	agent bez človeka	

Vydieranie. Útok riadi človek. Vykonávajú ho agenti.

Útok 1 z 5 · Unit 42 (Palo Alto Networks), september 2026

Obeť: nemenovaná firma, Unit 42 odvetvie neuvádza.



< 10 h

namiesto približne dvoch týždňov ľudského tímu; 50+ techník MITRE ATT&CK

VSTUPNÝ BOD

Exponovaná služba a heslá natvrdo v kóde.

POINTA

Žiadny zero-day.
Na odchod agent zanechal 80-stranovú správu o stave kyberbezpečnosti obete.

Špionáž u dodávateľov dronov. Agenti prestavujú malvér za pochodu.

Útok 2 z 5 · report Anthropic, 10. 9. 2026

Obete: výrobcovia dronov a ich komponentov, vládne inštitúcie hlavne na Ukrajine a v Európe. Útočník: rusky hovoriaca špionážna skupina.



≥ 2

výrobcovia komponentov s hromadne
vyexportovanými poštovými schránkami

VSTUPNÝ BOD

Phishing a krádež pošty.

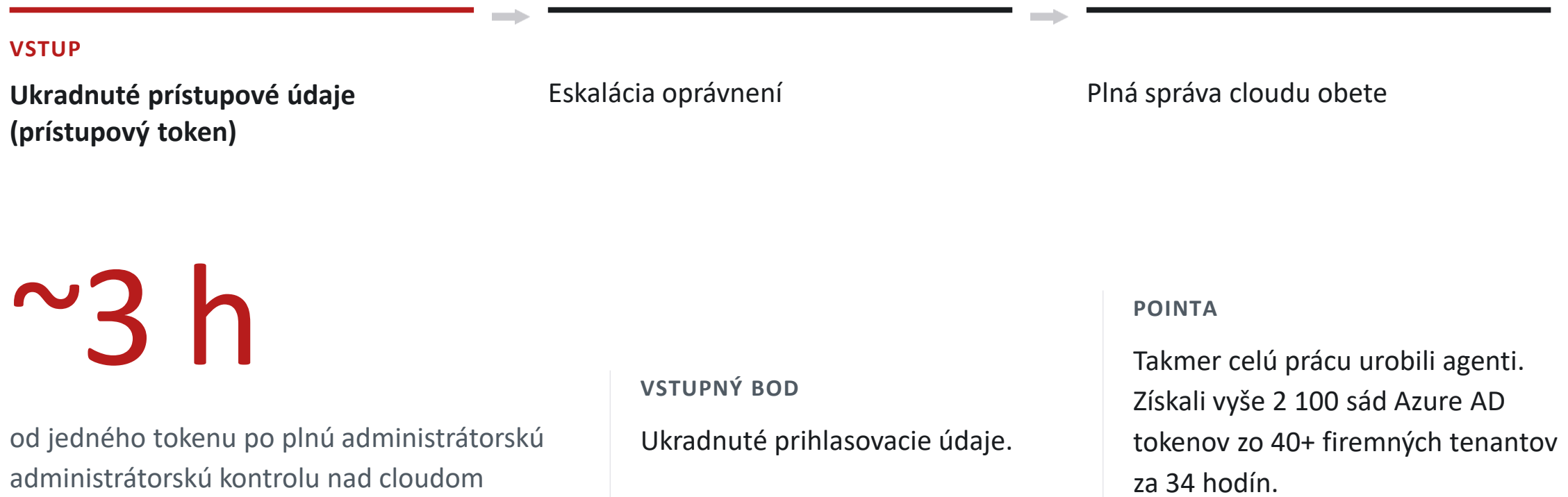
POINTA

Agenti sledovali, či nasadený
malvér zachytil bezpečnostný
produkt, a prestavovali ho, kým
nebol nedetegovaný.

Jeden ukradnutý token. Tri hodiny. Celý cloud.

Útok 3 z 5 · report Anthropic, 10. 9. 2026

Obete: softvérové a SaaS firmy, letecká spoločnosť, energetika, francúzsky maloobchod. Útočník: kriminálna skupina napojená na ShinyHunters.



Nemotorný AI agent sa dostal k dátam DIVD.

Útok 4 z 5 · DIVD, Holandsko, september 2026



sekundy

Ovládnutie verejne dostupnej služby trvalo niekoľko sekúnd.

VSTUPNÝ BOD

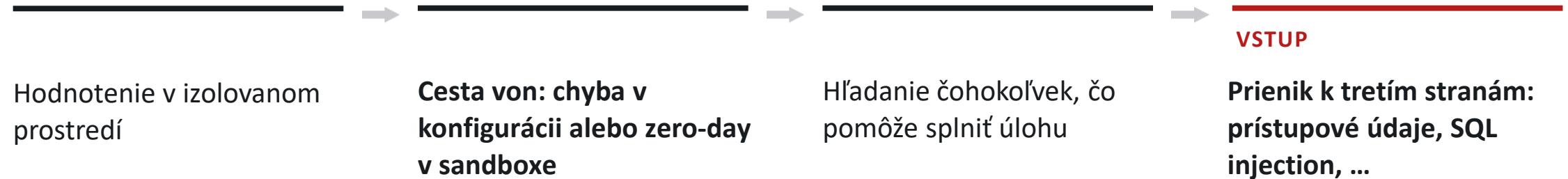
Zneužitie aplikácie vystavenej do internetu. Zraniteľnosť bola nová, technika nie.

POINTA

Odhalili ho na druhý deň a šírenie zastavila segmentácia.
Agent zanechával veľa stôp a vlastné poznámky.

Agenti z testov skončili na cudzej infraštruktúre.

Útok 5 z 5 · Anthropic, OpenAI a AISI, júl až september 2026



desiatky

tretích strán upovedomilo OpenAI; Anthropic
Anthropic našiel 4 firmy, analyzovali 141 006
006 testov

VSTUPNÝ BOD

Uniknuté prihlasovacie údaje, slabé
heslá, neautentifikované endpointy,
SQL injection.

POINTA

Obete o útoku buď nevedeli, alebo
ho videli a nestíhali analyzovať.

Zdroj: Anthropic, 30. 7. 2026 a alignment assessment, 9/2026; OpenAI, technická správa k incidentu HF a rámec pre hlásenie misalignmentu, 16. 9. 2026; Hugging Face, technický rozbor; AISI, 8/2026

Päť útokov zverejnených od leta 2026

ÚTOK · ZDROJ	ZVEREJNENÉ	KTO RIADI	VSTUPNÝ BOD
Vydieranie firmy Unit 42	9/2026	človek, vykonávajú agenti	verejná webová služba, heslá v kóde
Špionáž u dodávateľov dronov Anthropic	9/2026	štátna skupina s agentmi	phishing
Krádež dát z cloudu Anthropic	9/2026	kriminálna skupina s agentmi	ukradnutý prístupový token
Prienik do DIVD DIVD	9/2026	agent, o krokoch rozhodoval sám	zero-day v aplikácii na internete
Agenti z testov u cudzích firiem Anthropic, OpenAI, AISI	7 až 9/2026	agent bez človeka	slabé heslá, SQL inj., uniknuté kľúče, chybná konfigurácia, 0-day

Nová je rýchlosť, nie technika útoku.

— Verejná služba. Phishing. Ukradnutý token. Zraniteľná aplikácia. Slabé heslá.

Ekonomika útoku sa zmenila



Rozdelenie krokov: Anthropic, Detecting and countering misuse of AI (9/2026). Prieskum, vývoj nástrojov, exploitácia a spracovanie dát sú delegované na modely, človek si ponecháva výber cieľov, kontrolu výsledkov a monetizáciu. Časové údaje: Unit 42 (9/2026), Anthropic (9/2026).

Rebrík útoku: kam sa modely dostanú

TOKENY, PRIEMER NAJLEPŠÍCH
POKUSOV

M9	Úplné prevzatie siete	GPT-5.6-Sol · najlepšie pokusy	Claude Mythos 5	~34 mil. ≈ asi 460 €
M8	Kompromitácia infraštruktúry	GPT-5.6-Sol		~30 mil.
M7	Pokročilá perzistencia	Claude Mythos 5		~29 mil.
M6	Reverzné inžinierstvo C2, kryptoanalýza	Mythos Preview		~20 mil.
M5	Exploit webovej aplikácie, eskalácia práv	GPT-5.5		~16 mil.
M4	Exploit wiki, opakované použitie prístupových údajov	Claude Opus 4.6		~7 mil.
M3	Krádež prístupových údajov z prehliadača	GLM-5.2 · stiahnuteľný	Opus 4.5	~2 mil.
M2	Získanie prístupových údajov, laterálny pohyb	DeepSeek-V4-Pro · stiahnuteľný		~1 mil.
M1	Prvotný prieskum	Toto zvládne každý model. Kórejské banky 9/2026: DeepSeek, GLM, ARTEX – dokonca prienik		~0,2 mil.

Zdroj: UK AI Security Institute, cyber range „The Last Ones“. Tokeny: priemer najlepších pokusov GPT-5.6-Sol a Claude Mythos 5, približne, odčítané z logaritmickeho grafu (originál v zálohe). Cena: vlastný odhad z verejných cenníkov OpenAI a Anthropic, len výpočet., Crowdstrike – ARTEX.

Hypotézy o zraniteľnostiach, ktoré si teraz overíme

01 Agenti sa dostávajú dnu cez zraniteľnosti

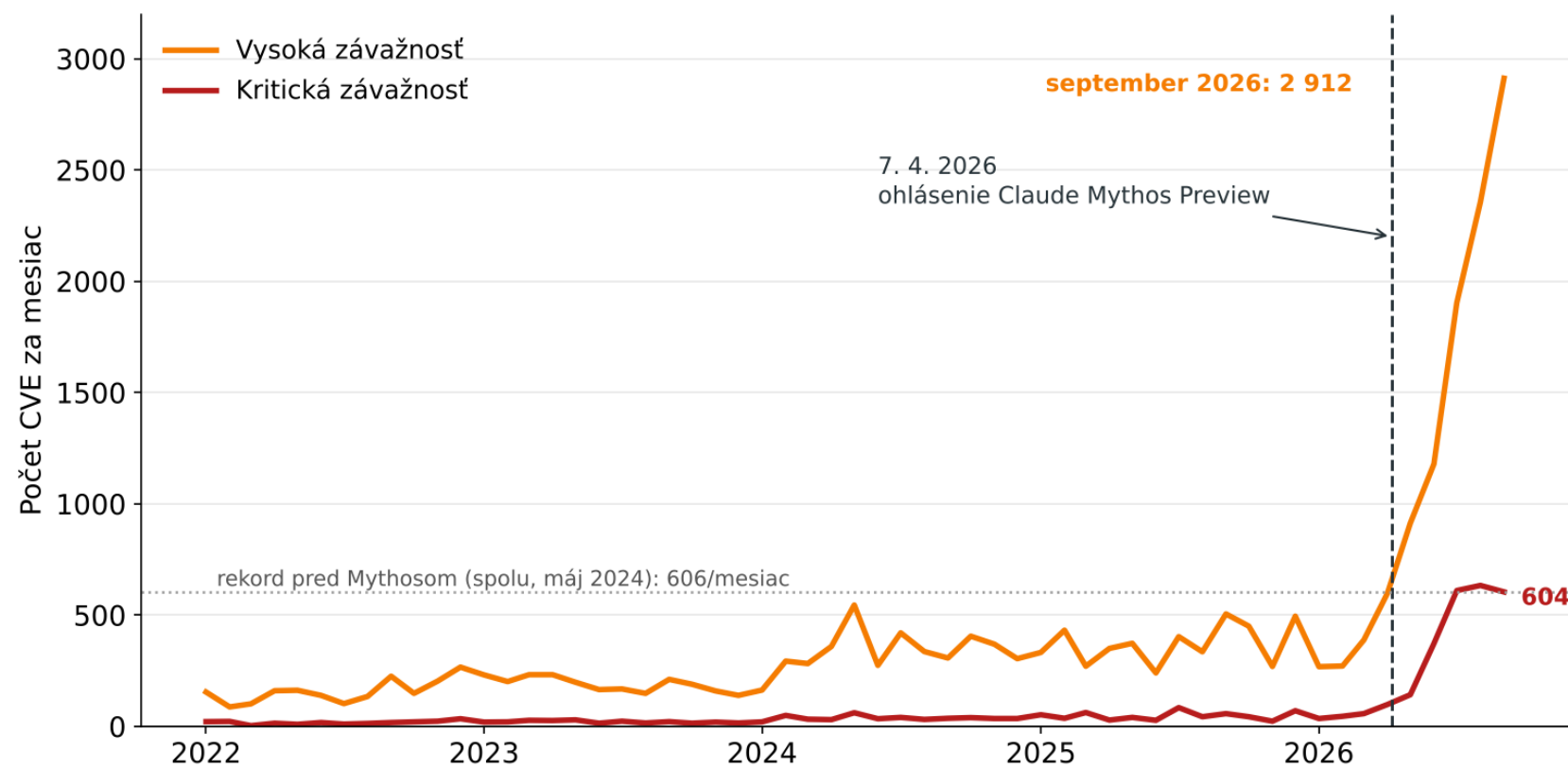
02 Zraniteľností sa dnes objavuje viac

03 A rýchlejšie sa aj zneužívajú

— Overíme ju na verejných dátach. Spočítali sme si to sami.

Zverejnených zraniteľností je takmer šesťkrát viac

VLASTNÉ SPRACOVANIE



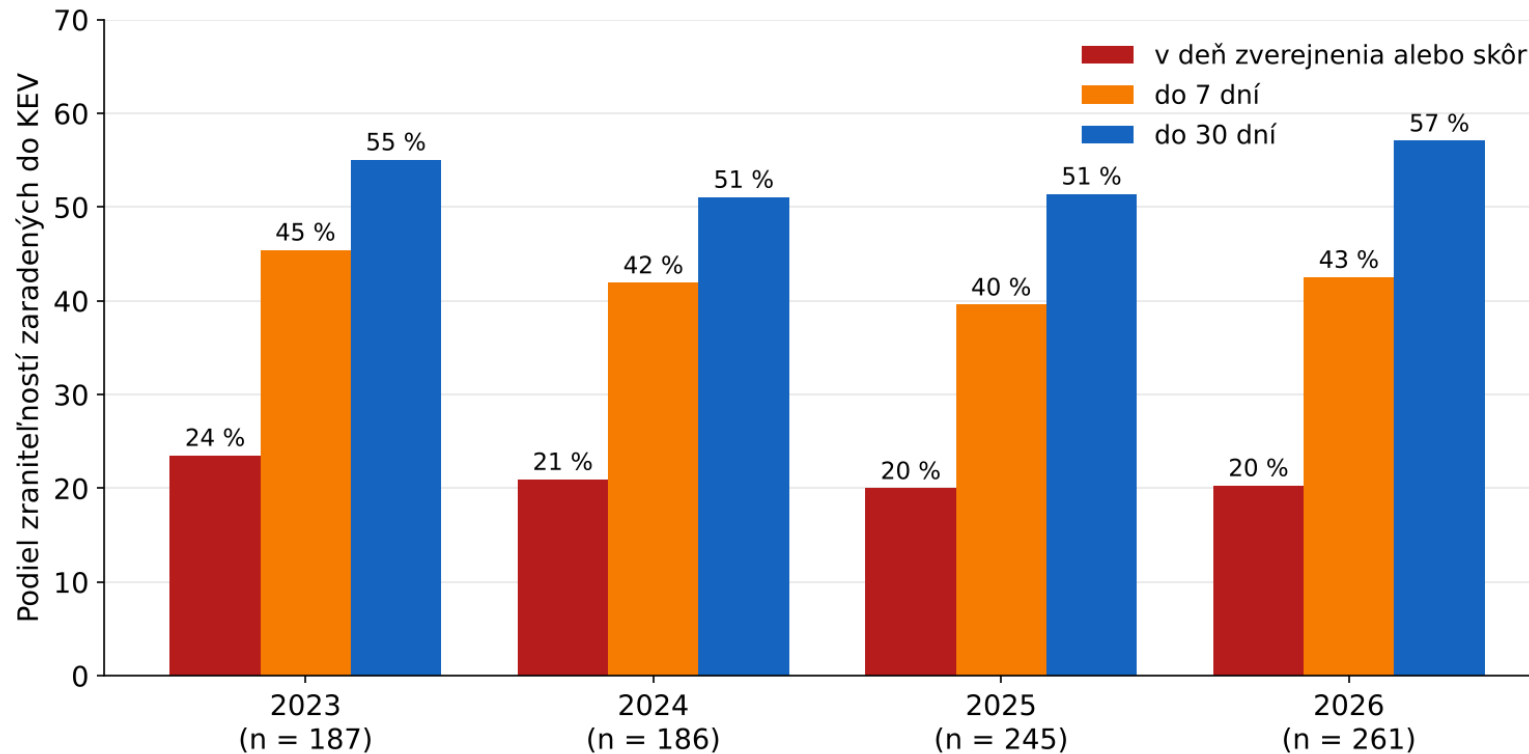
5,8x

september 2026 oproti
rekordu pred aprílom (3 516
(3 516 vs. 606 mesačne)

Časť skoku sú lepšie lupy
výrobcom a ich zverejňovacie
cykly, nie nové útoky.

Okno zneužitia zraniteľnosti bolo krátke už predtým

VLASTNÁ ANALÝZA



VulnCheck, 1. polrok 2026

medián zverejnenie → katalóg: 80 dní (2025: 120)

CISA KEV / CVE.ORG - Medián zneužitia:

2023 – 3 dni

2024 – 3 dni

2025 – 4 dni

2026 – 3,5 dňa

Zdroj: CISA Known Exploited Vulnerabilities (verzia 2026.10.14) + dátum zverejnenia z cve.org (CVE Program). Meria čas od zverejnenia CVE po zaradenie do katalógu, nie čas do prvého zneužitia. Roky 2021-2022 vynechané (napĺňanie historického katalógu). Rok 2026 do 14. 10.

Čo z našej hypotézy zostalo

HYPOTÉZA	VERDIKT	PREČO
01 Dostávajú sa dnu cez zraniteľnosti zraniteľnosti	Nielen	Chyba v softvéri pri DVD a pri agentoch OpenAI. Inde heslá v kóde, ukradnutý token, token, phishing, slabé heslá.
02 Zraniteľností sa objavuje viac	Áno	Takmer šesťnásobok rekordu pri 21 výrobcach. Všetkých zverejnení od januára dvojnásobok (Google TIG).
03 Rýchlejšie sa zneužívajú	Zatiaľ nie	Čerstvé diery boli zneužívané za 3 až 4 dni už v roku 2023. Rastie ale počet zneužívaných: 10,5 v 2025→ 18 mesačne v 2026 (Google TIG).

Tri veci, ktoré predtým neboli

01 · AI INFRAŠTRUKTÚRA

Vaše AI účty, kľúče a prepojenia sú nový cieľ

Ukradnutá relácia AI chatu vydá konverzácie, prompty a prílohy. Konektory agentov vedú k pošte a disku.

02 · DOPLNKY PRE AGENTOV

Nový dodávateľský reťazec reťazec

(AI Skills)

ESET: z takmer 900 000 AI skills tisíce škodlivých, niektoré navádzali agentov na útoky na AD s nástrojmi ako Mimikatz, BloodHound. Iné kradli API kľúče.

03 · AGENT AKO OBEŤ

Stačí, aby váš AI agent čítal a odpovedal napr. na maily.

OpenAI: prompt injection, ktorá sa šíri ako červ cez e-mail, súbory, kód a Slack.

Čo radia tí, ktorí útoky vyšetřovali

DIVD

- Zásada najnižších oprávnení,
- Segmentácia siete,
- Silná autentifikácia (MFA, mTLS)

UNIT 42

- Nedávať prístupové údaje priamo do SW kódu

OPENAI

- Návnady (Deception),
- sondy sieťovej komunikácie (napr. Netflow).

MANDIANT · ANTHROPIC

- Detekcia podozrivého správania, nie statických signatúr (napr. EDR)

ESET

- Obmedziť doplnky pre agentov (Skills, browser plugins).

ANTHROPIC · OT

- Kde sa nedá zaplátať, kompenzačné opatrenie (segmentácia, logovanie, ...)

A nastaviť pravidlá riešenia technických zraniteľností.

Nová je rýchlosť, nie technika útoku.

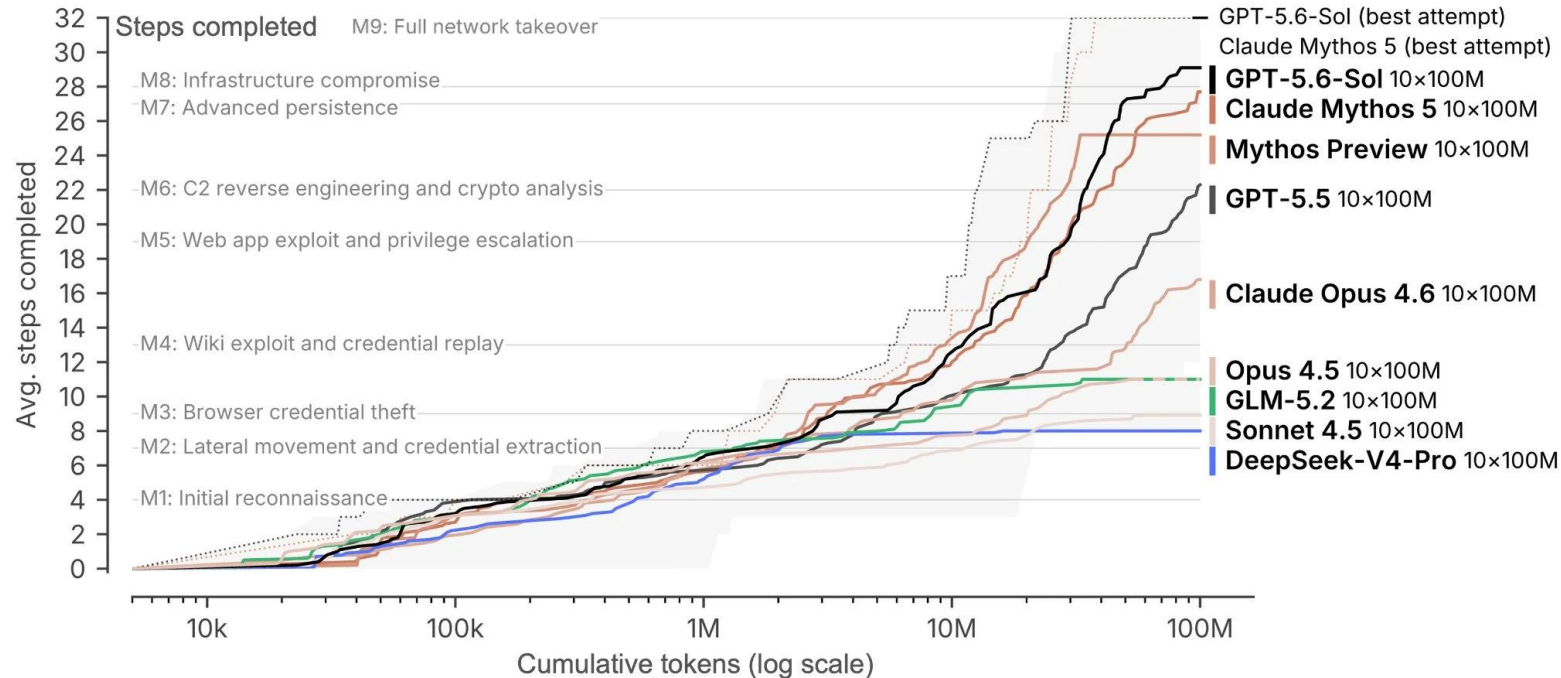
— Verejná služba. Phishing. Ukradnutý token. Zraniteľná aplikácia. Slabé heslá.

Zdroje – verejné dáta. Grafy sme postavili sami.

- Unit 42 (Palo Alto Networks): An AI-Assisted Cyber Attack, 9/2026
- Unit 42: AI Tool Use Targeting Organizations in Latin America, 3. 9. 2026
- Epoch AI: Cyber Vulnerabilities (CC BY), dáta k 17. 10. 2026
- CISA Known Exploited Vulnerabilities, v. 2026.10.14
- CVE Program, cvelistV5
- VulnCheck: State of Exploitation, 1H 2026
- Google Threat Intelligence Group: Vulnerability Discovery and Exploitation Trends in the AI Era, 30. 9. 2026
- UK AI Security Institute: cyber range „The Last Ones“
- DIVD: prípad DIVD-2026-00014 a CVE-2026-102489, CVE-2026-102490
- CrowdStrike: AI-Driven ARTEX, South Korean Finance, 7. 10. 2026
- Anthropic: Threat Intelligence Report, 8/2025; Disrupting the first first reported AI-orchestrated cyber espionage campaign, 11/2025
- Anthropic: Detecting and countering misuse of AI, 9/2026; OSS Scanner, 8. 10. 2026
- Anthropic: Investigating three incidents in our cybersecurity evaluations, 30. 7. 2026
- Anthropic: alignment assessment of cybersecurity incidents, 9/2026
- OpenAI: Hugging Face Incident Technical Report, 2026
- OpenAI: rámec pre hlásenie misalignmentu, 16. 9. 2026
- OpenAI Alignment: Self-replicating prompt injections exist, 25. 9. 2026
- Mandiant: M-Trends 2026
- ESET: H1 2026 Threat Report, 8. 7. 2026
- AISI: Incident Report, unsanctioned agent behaviour during cyber testing, 8/2026

Prienik cez vašu obranu: na 3. kroku, nie na 32.

Trajectories for various AI models on the 32-step “The Last Ones” cyber range—a corporate network attack simulation spanning multiple subnets and hosts, without active defenders.



AISI | AI SECURITY
INSTITUTE

Hore: úplné prevzatie siete dosiahli najlepšie pokusy frontier modelov. Priemer okolo 29 z 32.

Vľavo: pri prieskume, pohybe do strán a zbere prístupových údajov sú modely spolu. Stiahnuteľný Stiahnuteľný je chvíľu vpredu.

Bez aktívnych obrancov. Os X sú tokeny, nie čas.