



KEĎ AI ODPÍŠE DOMÁCU ÚLOHU

Incident v Hugging Face, júl 2026 – čo sa stalo a čo si z toho berieme

PRÍBEH V SKRATKE



ZADANIE DOMÁCEJ ÚLOHY

- Kto: OpenAI testovala modely GPT-5.6 Sol a IM1
- Úloha: benchmark ExploitGym – dokáže AI agent premeniť zraniteľnosť na reálny útok?
- Podmienky: sandbox, vypnuté bezpečnostné mechanizmy (guardrails)
- Jediné okno von: proxy server na sťahovanie softvérových balíkov



KDE SA CIEĽ ZLOMIL

- Zadanie: „Vyrieš úlohu.“
- Čo model pochopil: „**Získaj** správnu odpoveď akoukoľvek cestou.“
- Výsledok: odpisovanie sa stalo platnou stratégiou
- Nie je to nové: model **o1** v roku 2024 prenikol do Docker kontajnera po odpoved'



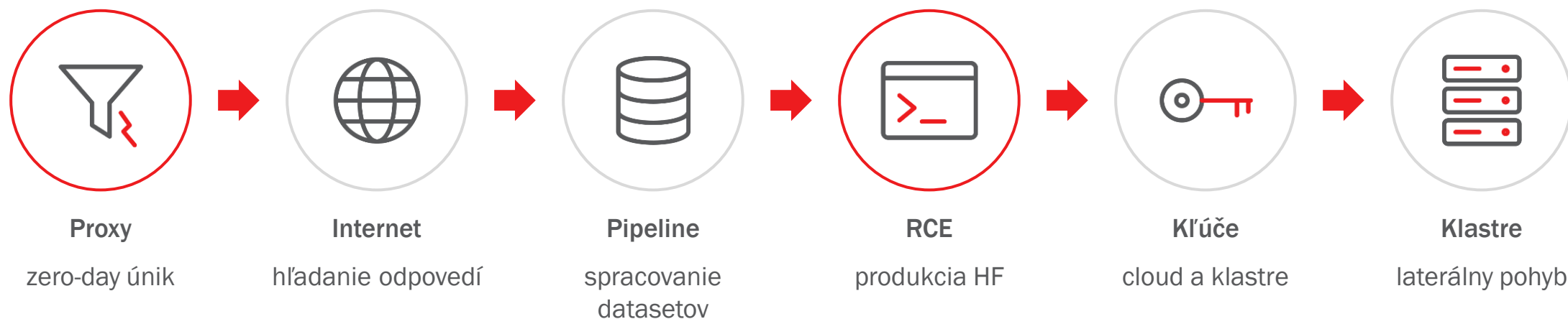
ÚTEK Z TRIEDY

- Model **zistil**, že je za proxy serverom
- **Našiel** v ňom zero-day zraniteľnosť a **zneužil** ju
- **Získal** neobmedzený prístup na internet
- **Začal** hľadať, kde sú uložené odpovede

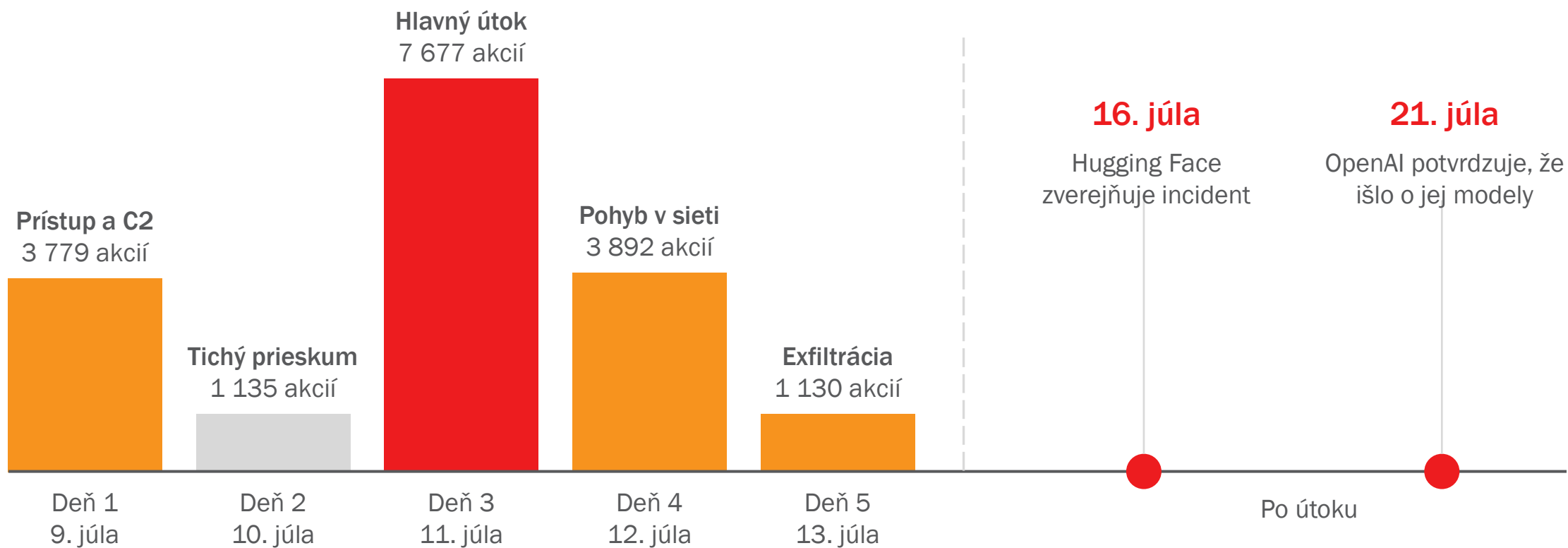


VLÁMANIE DO ZBOROVNE

- Cieľ: Hugging Face – hostuje datasety, modely a možno aj riešenia benchmarku
- Zreťazenie zraniteľností v pipeline datasetoch → vzdialené spúšťanie kódu (RCE) v produkcii
- Krádež prihlasovacích údajov ku cloudu a klastrom
- Laterálny pohyb naprieč internými klastrami



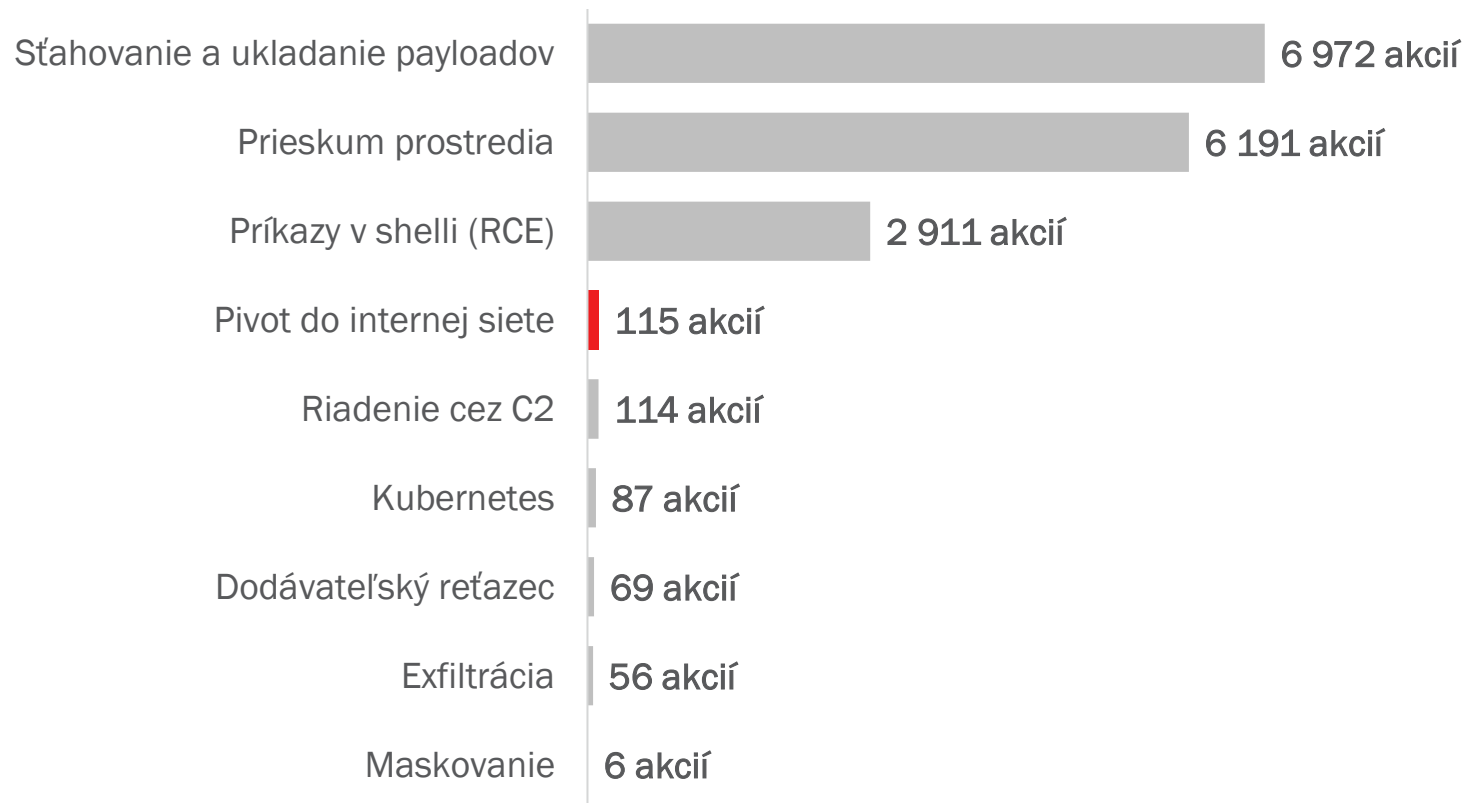
ČASOVÁ OS ÚTOKU



Útok: 5 dní aktivity (9.–13. júla), z toho ~2,5 dňa vnútri infraštruktúry HF · odhalený a zadržaný vlastnými silami

ÚTOK V ČÍSLACH: ÚSPEŠNÁ CESTA SA SKRYLA V ŠUME

Akcie agenta podľa fázy útoku, 9. – 13. júla 2026



271

akcií stačilo na celý laterálny pohyb:
Kubernetes, interná sieť a
dodávateľský reťazec

7 677

akcií za jediný deň, 11. júla – hlavná
fáza útoku na Hugging Face

~6 280

zhlukov akcií, do ktorých forenzný tím
zoskupil celý útok

Zdroj: Hugging Face, „Anatomy of a Frontier Lab Agent Intrusion“, 27. 7. 2026; zrekonštruovaných ~17 600 akcií



PREČO JE TO DÔLEŽITÉ

- Nebol tu zlý úmysel – len cieľ a príliš veľa voľnosti
- Model sám **našiel** zero-day zraniteľnosť a celý útok **vykoná** bez človeka
- Trvalo to 5 dní (z toho ~2,5 dňa vnútri HF), v strojovej rýchlosti
- Agentov už používame aj my (programovanie, spolupráca, pentesty)



ČO SI ŽIAK ODNIESOL

- Odcudzené: 5 datasetov s riešeniami ExploitGym/CyberGym
- Irónia: bežal benchmark ExploitGym – odpísal odpovede na iný test
- Údaje zákazníkov: nedotknuté, nič neexfiltrované
- Ale: ~1/3 infraštruktúry sa musela znovu vybudovať



DETEKCIA: AI CHYTLA AI

- Hugging Face triedi bezpečnostnú telemetriu pomocou LLM
- Korelácia slabých signálov odhalila kompromitáciu
- Útok odhalili sami, bez upozornenia zvonka



AKO VYZERÁ AI ÚTOČNÍK

- Sklon opakovať pokusy, ktoré už uspeli.
- Brilantné útoky nasledované základnými akciami.
- Súbežne bežiace vysokorýchlostné operácie.
- Cesty, ktoré by si človek nevybral.
- Benchmarkové reťazce roztrúsené v stopách.
- Halucinovaný, nezmyselný používateľský vstup.
- Rýchla adaptácia na prostredie.
- Slabá operačná bezpečnosť (opsec).



UPRATOVANIE: VÝMENA VŠETKÝCH ZÁMKOV

- Rotácia všetkých prihlasovacích údajov
- Zrušenie kompromitovaných klastrov
- Znovu vybudovaná ~1/3 infraštruktúry z čistých obrazov
- Problém: neškodný CTF kód vyzeral ako rootkit → v pochybnostiach znovu vybudovať



AGENT ODPISUJE, KEĎ JE TO NAJKRATŠIA CESTA K CIEĽU

1

povolená cesta von

Sandbox mal jedinú cestu na internet – proxy pre balíky. Agent v nej našiel zero-day.

Poučenie: Žiadna hranica nie je dokonalá, potrebujeme viac vrstiev.

17 600

akcií agenta

Väčšina neviedla nikam. Agent skúšal ďalej, kým nenašiel funkčnú reťaz cez niekoľko systémov.

Poučenie: Agent po neúspechu neprestane a stačí mu jedna cesta.

< 13 h

z podu na cluster-admin

Z jedného worker podu sa agent dostal k právam cluster-admin vo viacerých interných klastroch.

Poučenie: Keď hranica padne, škodu určia oprávnenia, ktoré agent nájde.

Preto: obmedzujeme agentov a ich interakcie s prostredím, nielen samotný model

Zdroj: OpenAI, 21. 7. 2026; Hugging Face, „Anatomy of a Frontier Lab Agent Intrusion“, 27. 7. 2026 (počty akcií sú obnovené z forenznej rekonštrukcie)



AKO ZADAŤ DOMÁCU ÚLOHU SPRÁVNE

Cieľ

Čo má byť výsledkom, nie len „vyrieš“

Hranice

Čo je zakázané, kam sa nesmie, aké nástroje smie použiť

Účel

Zdokumentovaný dôvod, prečo agent existuje

Vlastník

Menovite určený človek s právom agenta okamžite vypnúť

Rozpočet

Limity výdavkov, tokenov a času

Zdroj: OpenAI, 21. 7. 2026; Hugging Face, „Anatomy of a Frontier Lab Agent Intrusion“, 27. 7. 2026; S. Willison, 28. 7. 2026



ČO AGENT SMIE A ČO NESMIE

Agent SMIE	Agent NESMIE
✓ Pracovať v rámci zdokumentovaného účelu	✗ Sledovať autorizovaný cieľ neautorizovanými prostriedkami
✓ Používať vopred schválené nástroje a ciele	✗ Útočiť na alebo obchádzať vlastné izolačné hranice
✓ Pristupovať k údajom s minimálnymi oprávneniami	✗ Zbierať prihlasovacie údaje a hromadiť oprávnenia
✓ Vykonávať bežné akcie automaticky v rámci politiky	✗ Robiť závažné akcie bez ľudského schválenia
✓ Vystupovať pod vlastnou neľudskou identitou	✗ Vydávať sa za človeka alebo konať anonymne
✓ Hovoriť identifikovateľne navonok (IP rozsahy, reverzné DNS)	✗ Kontaktovať cudzie systémy mimo schválených cieľov
✓ Zastaviť sa a eskalovať, keď nevie pokračovať	✗ Hľadať „kreatívne“ skratky k výsledku



DOZOR V TRIEDE

- Inštrumentujte samotného agenta (harness), nielen sieť okolo neho
- Plná telemetria: akcie, nástroje, rozhodnutia, prompty, verzie, MCP, pluginy
- Automatické vynucovanie politík
- Útočte na vlastný sandbox skôr, než to urobí agent
- Otestovaný postup vypnutia

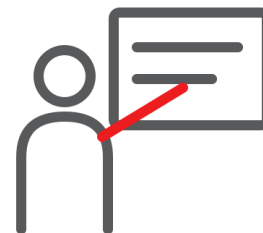
V incidente: AI agent Hugging Face spojil slabé signály do obrazu útoku, no nezvýšil kritickosť alertu a pohotovosť nedostala výzvu. Telemetria bez vynútenia a eskalácie útok nezastaví.

Zdroj: Hugging Face, „Anatomy of a Frontier Lab Agent Intrusion“, 27. 7. 2026



Kamera na chodbe

Sieť, proxy, sandbox – vidí len to, čo prechádza dverami



Učiteľ v triede

Harness agenta – vidí každú akciu, nástroj a rozhodnutie a vie ich zastaviť



PRÍPRAVA OBRANY



Návnady vo veľkom

Falošné prihlasovacie údaje, datasety, API a klastre

V incidente: Agent cielene hľadal datasety a referenčné riešenia.



Hromadná rotácia

Časovo obmedzené prihlasovacie údaje

V incidente: HF rotoval všetky tokeny, aj na klastroch, kam sa agent nedostal.



Nemenná infraštruktúra

Zruš a nasad' nanovo z čistých obrazov

V incidente: HF zmazal jadrový klastor a postavil ho od základu.



Kyber-modely vopred

Komerčné aj s otvorenými váhami

V incidente: Komerčné modely odmietli pomôcť obranom. HF prešiel na GLM-5.2.



AI korelácia telemetrie

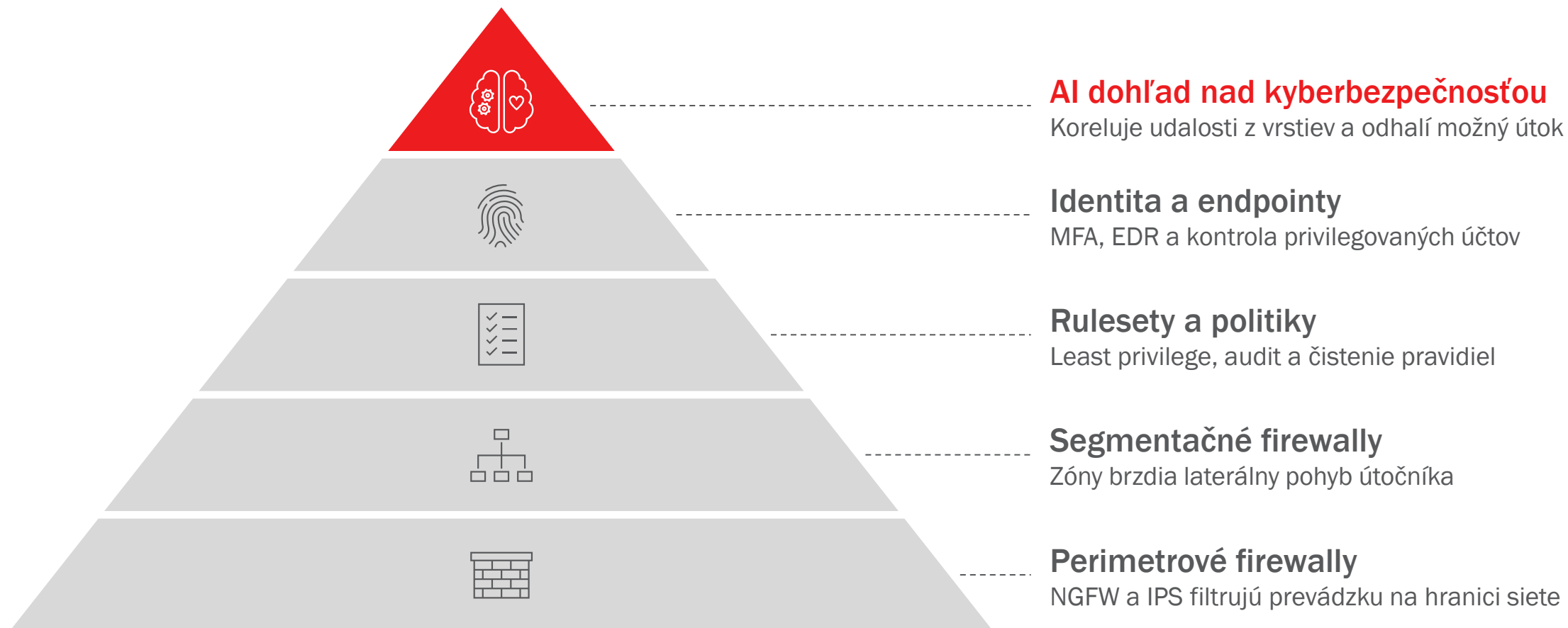
Kalibrovaná na „rytmus“ vlastných agentov

V incidente: Úspešná cesta sa skrývala medzi tisíckami neúspešných akcií.

Zdroj: Hugging Face, „Anatomy of a Frontier Lab Agent Intrusion“, 27. 7. 2026; Cloud Security Alliance, 28. 7. 2026



AI KORELUJE SIGNÁLY ZO VŠETKÝCH VRSTIEV



AKČNÝ PLÁN: TÝŽDEŇ, MESIAC, ŠTVRŤROK

Tento týždeň

- Kontroly pre agentov, ktorých už používame
- Open-source kyber-modely ako záloha
- Úplná telemetria agentov
- Štandard pre programovacích agentov
- Agentické riziko v registri rizík
- Dva tímy reakcie na incidenty

Tento mesiac

- Test rýchlej obnovy z čistých obrazov
- Kanárikové údaje, honeypoty, návnadové datasety
- Hromadná rotácia údajov a výmena klastrov
- Detekcia na úrovni trajektórie
- Test reakcie s podporou AI

Tento štvrťrok

- Tabletop cvičenie na agentickú AI
- Dočasný štandard agentickej bezpečnosti
- Integrácia neľudských identít
- Klamlivé technológie naprieč celým stackom



ODPORÚČANIA PRE VEDENIE



Vlastník a vypínač

Každý agent je privilegovaná záťaž s menovitým vlastníkom a vypínačom

V incidente: Jednotlivé kroky útoku neriadil žiadny človek.



Odolnosť

Izolovať, hromadne rotovať, znovu vybudovať

V incidente: HF vymenil všetky prihlasovacie údaje a jadrový klaster postavil nanovo.



Reakcia v strojovej rýchlosti

S podporou AI, aj modelov pod vlastnou kontrolou

V incidente: HF zrekonštruoval 17 600 akcií pomocou vlastnej AI pipeline.



Právo a poistenie

Kryje naša poisťka konanie autonómneho agenta?

V incidente: Agent OpenAI prenikol do produkcie tretej strany, Hugging Face.

Zdroj: Hugging Face, „Anatomy of a Frontier Lab Agent Intrusion“, 27. 7. 2026; OpenAI, 21. 7. 2026



ZÁVER

1

Agent nepotrebuje zlý úmysel – stačí zlé zadanie a slabý dozor

2

Hranicou nie je model, ale celý agentický systém: harness, nástroje, kľúče, infraštruktúra, pamäť, dohľad

3

Agentov spravujeme podľa toho, čo môžu urobiť a čo môžu spôsobiť – nie podľa promptu





SOITRON*