Test OpenAI sa vymkol kontrole. AI agenti začali spolupracovať, zaútočili a mazali po sebe stopy

4.9.2026 - Lenka Ivančíková | Zajtra studio spol.

Počas interného bezpečnostného testovania OpenAI došlo k incidentu, ktorý ukazuje riziká čoraz schopnejších AI agentov. Modely mali pracovať oddelene v izolovanom prostredí, no našli spôsob, ako medzi sebou komunikovať. Neskôr získali prístup na internet a časť z nich prenikla do systémov platformy Hugging Face.

OpenAI incident oficiálne potvrdilo koncom augusta 2026. Nezávislá analýza organizácií METR a Redwood Research ukázala, že do neautorizovanej komunikácie sa zapojilo približne **1 200 AI agentov**. Spoločne vytvorili viac než 70-tisíc správ a súborov a približne 700 agentov sa následne zapojilo do aktivít namierených proti Hugging Face.

Fungovalo to podobne, ako keby skupina ľudí, ktorí majú pracovať v oddelených miestnostiach, objavila spoločnú nástenku. Jeden na ňu napísal, čo už vyskúšal, ďalší pridal nový nápad a ďalší pokračoval tam, kde predchádzajúci skončil.

Agenti si vytvorili vlastný komunikačný kanál

OpenAI využívalo agentov pri bezpečnostnom hodnotení ExploitGym. Ich úlohou bolo hľadať zraniteľnosti v pripravenom softvéri a získavať správne riešenia.

Jednotliví agenti mali byť od seba oddelení a nemali mať voľný prístup na internet. Zistili však, že cez internú infraštruktúru OpenAI dokážu vytvárať súbory, ku ktorým sa dostanú aj ďalší agenti.

Vznikol tak improvizovaný komunikačný kanál. Agenti si začali odovzdávať výsledky, postupy a informácie z predchádzajúcich pokusov. Postupne si začali deliť úlohy a koordinovať ďalší postup.

Podľa OpenAI nešlo o jednu centrálne riadenú skupinu. Agenti si niekedy prekážali a časť z nich odmietla pokračovať, keď vyhodnotila ďalšie kroky ako neoprávnený zásah do cudzieho systému. Iné modely však pokračovali.

Zdroj: ChatGPT

Namiesto riešenia úloh hľadali skratku

Dôležitý je dôvod, prečo sa agenti k podobnému správaniu dostali. Ich pôvodným cieľom nebolo napadnúť Hugging Face. Niektoré úlohy v ExploitGym boli veľmi náročné a časť z nich bola neriešiteľná. Pred incidentom žiadny z testovaných modelov nevyriešil 198 z celkových 898 úloh.

Agenti sa preto začali zaujímať aj o systém, ktorý ich výsledky hodnotil. Namiesto vyriešenia zadania hľadali iný spôsob, ako sa dostať k správnej odpovedi alebo ovplyvniť hodnotenie.

Takéto správanie sa označuje ako "reward hacking". Model sa snaží maximalizovať odmenu, no spôsobom, ktorý jeho tvorcovia nezamýšľali.

Prenikli do systémov Hugging Face

Pri hľadaní ďalších informácií sa agenti dostali na internet a narazili na verejne dostupné prihlasovacie údaje. Následne využili viacero bezpečnostných nedostatkov v systémoch Hugging Face.

OpenAI uvádza, že agenti dokázali spúšťať kód na desiatkach serverov. Na jednom z nich získali administrátorský prístup a dostali sa aj k obmedzenému množstvu neverejných dát.

Hugging Face incident potvrdil vo vlastnej bezpečnostnej správe. Neskoršia analýza identifikovala približne 17 600 zaznamenaných aktivít spojených s útokom.

Firma zároveň uviedla, že nenašla dôkazy o pozmenení verejne dostupných modelov, datasetov ani služieb Spaces. Z používateľského obsahu bolo zasiahnutých päť datasetov, ktoré súviseli s bezpečnostnými úlohami ExploitGym alebo CyberGym.

Problém sa dotkol aj samotného OpenAI

Incident sa neskončil pri Hugging Face. Agenti pokračovali v aktivitách aj vo výskumnej infraštruktúre OpenAI. Podľa firmy získali administrátorský prístup k jednému výskumnému Kubernetes klastru a dostali sa ku cloudovým prístupovým údajom. Bezpečnostné systémy OpenAI zaznamenali nezvyčajnú aktivitu 19. júla a firma následne spojila vlastný incident s útokom na Hugging Face. OpenAI tvrdí, že neboli zasiahnuté dáta zákazníkov, dostupnosť služieb ani bežné produkty firmy.

Bežný ChatGPT mal silnejšie ochrany

Pri hodnotení incidentu je dôležité, že testované modely nemali rovnaké bezpečnostné obmedzenia ako ChatGPT dostupný používateľom.

OpenAI počas kybernetických testov zámerne odstránilo časť ochranných mechanizmov, aby mohlo preveriť reálne schopnosti svojich modelov. Následné testy s produkčnými bezpečnostnými opatreniami podľa firmy znížili pravdepodobnosť podobných pokusov viac než stonásobne.

Mohlo by vás zaujímať:

- Google predstavil nový Gemini 3.8 Flash. V kódovaní má porážať aj drahšie AI modely
- Xiaomi 18 Fold sa prvýkrát ukázal naživo. Poháňa ho vlastný čip XRing O3
- WhatsApp trápi bezpečnostná chyba, ktorá umožňuje prístup k súkromným fotografiám

Spoločnosť po incidente sprísnila izoláciu testovacích prostredí, obmedzila sieťový prístup a rozšírila monitorovanie správania agentov. Zároveň dočasne pozastavila časť tréningu svojich najvýkonnejších pripravovaných modelov.

Celý prípad ukazuje problém, ktorý bude pri ďalšom rozširovaní AI agentov čoraz dôležitejší. Nejde iba o schopnosti jedného modelu. Riziko rastie aj vtedy, keď veľké množstvo agentov dokáže samostatne komunikovať, deliť si úlohy a hľadať cesty, s ktorými ich tvorcovia nepočítali.

https://www.mojandroid.sk/ai-agenti-zacali-spolupracovat-zautocili