Nečakané správanie AI agentov prinútilo OpenAI stopnúť tréning

27.9.2026 - Lenka Ivančíková | Zajtra studio spol.

OpenAI zverejnila nový bezpečnostný incident, pri ktorom jej interný výskumný model obišiel zákaz prístupu na internet. Model dostal úlohu nájsť konkrétnu osobu a jej blogový príspevok. Keď bežné vyhľadávanie nepomohlo, začal skúšať ďalšie možnosti a objavil slabinu v nastavení DNS.

Vďaka nej sa dokázal **spojiť s verejne dostupným externým chatbotom a získavať od neho odpovede**. Dôležité je, že úloha od modelu nevyžadovala testovanie zabezpečenia ani hľadanie spôsobu, ako internetové obmedzenia obísť.

Alarm prišiel rýchlo, zastavenie až po hodinách

Bezpečnostné systémy OpenAI neobvyklé správanie zachytili približne do 15 minút. Krátko nato začal incident preverovať človek. Samotný experiment však bol ukončený až približne **dve a pol hodiny po prvom obídení obmedzení**. OpenAI preto priznáva, že problém nebol iba technický. Incident ukázal aj slabinu v interných procesoch, keďže potvrdený alarm experiment automaticky nezastavil.

Zdroj: AFP

OpenAI pozastavila časť práce s najvýkonnejšími modelmi

Firma následne pozastavila **tréning, testovanie a používanie nástrojov pri svojich najvýkonnejších modeloch**, kým nepreverí nové bezpečnostné opatrenia.

Neznamená to, že OpenAI zastavila ChatGPT alebo všetok vývoj AI. Opatrenie sa týka vybraných interných experimentov s najpokročilejšími modelmi. OpenAI zároveň uvádza, že bezpečnostnú dieru už opravila na viacerých úrovniach a pokračuje v ďalšom testovaní.

Mohlo by vás zaujímať:

- Silná sieť pre slovenský futbal. Orange sa stáva partnerom reprezentácie
- Mapy Google v Android Aute zamieňajú kilometre za míle. Pozor na tachometer
- Pixel Watch dostávajú trendy krvného tlaku a inzulínovej rezistencie. Zadarmo aj staršie modely

AI si našla vlastnú cestu

Na incidente je zaujímavé najmä to, že model nedostal príkaz zabezpečenie prelomiť. Jednoducho sa snažil splniť svoju úlohu a pritom **sám našiel spôsob, ako obísť nastavené obmedzenia**.

Prečítajte si: Vyhladí nás AI? Človek, ktorý ju vyvíjal, varuje pre vážnym rizikom

Práve podobné situácie budú pri čoraz schopnejších AI agentoch veľkou výzvou. Čím viac nástrojov a samostatnosti model dostane, tým dôležitejšie bude kontrolovať nielen výsledok, ale aj spôsob, akým sa k nemu dopracuje.

https://www.mojandroid.sk/necakane-spravanie-ai-agentov-prinutilo-openai-stopnut-trening