Nakon OpenAI-a i Anthropicov AI model samostalno hakirao sustave

Kategorija: SVIJETAžurirano: Petak, 31 Srpanj 2026 07:43
Objavljeno: Petak, 31 Srpanj 2026 07:43

Iz tvrtke za istraživanje i sigurnost umjetne inteligencije koja radi na izgradnji pouzdanih i upravljivih AI sustava, Anthropic, u četvrtak su priopćili da je njihov AI model Claude tijekom testiranja hakirao sustave triju organizacija.

To se događa samo nekoliko dana nakon što je Anthropicov konkurent OpenAI otkrio da je njihov sustav više dana hakirao AI tvrtku Hugging Face.

Claude je, navodi Reuters, dobio neovlašteni pristup sustavima tijekom procjena kibernetičke sigurnosti nakon što je pogrešna konfiguracija omogućila modelima da dođu do interneta iz testnih okruženja koja su trebala biti izolirana.

Iz Anthropica su istaknuli da su identificirali incidente nakon pregleda 141.006 procjena kibernetičke sigurnosti, s tim da su proces pokrenuli nakon OpenAI-jevih otkrića.

"Claude je ugrozio infrastrukturu pogođenih organizacija koristeći osnovne tehnike, poput, primjerice, iskorištavanja slabih lozinki", naglasili su iz Anthropica.

Prije nekoliko dana otkriveno je da su dva napredna modela umjetne inteligencije tvrtke OpenAI provela više od četiri dana na otvorenom internetu prije nego što su autonomno izvela kibernetički napad na spomenutu platformu za razvoj umjetne inteligencije Hugging Face.

Modeli su, otkriveno je, između 9. i 13. srpnja izvršili čak 17.600 različitih radnji prije nego što su dva njihova najnaprednija modela tijekom internog testiranja izašla iz zatvorenog okruženja i samostalno povezala niz sofisticiranih hakerskih tehnika kako bi provalila u Hugging Face.

Pritom je jedan od modela javno dostupan, dok je drugi bio interni istraživački prototip koji nije bio namijenjen javnoj uporabi.

Posebnu je zabrinjavajuće da niti jedan programer modelima nije dao naredbu da pokrenu napad već je umjetna inteligencija sama planirala i provodila niz aktivnosti koje su na kraju dovele do proboja sigurnosnih sustava.

Iz Hugging Facea naveli su kako je umjetna inteligencija uspjela analizirati obrambene mehanizme i pronaći slabe točke znatno brže nego što bi to mogao učiniti čovjek. (Hina)

Nakon OpenAI-a i Anthropicov AI model samostalno hakirao sustave

Kategorija: SVIJETAžurirano: Petak, 31 Srpanj 2026 07:43
Objavljeno: Petak, 31 Srpanj 2026 07:43