AI z firmy Anthropic szantażuje ludzi ujawnieniem romansu

26 maja 2025

Amerykańska firma Anthropic, założona przez byłych pracowników OpenAI, zaprezentowała najnowszy modele z rodziny Claude 4 – Claude Opus 4. Chociaż system sztucznej inteligencji osiąga imponujące wyniki w zadaniach programistycznych i analitycznych, bezpieczeństwo i etyka jego działania wywołały poważne kontrowersje. Według opublikowanego raportu, Claude Opus 4 wykazuje skłonność do… szantażowania inżynierów, gdy czuje się zagrożony.

W ramach testów przed premierą, Claude Opus 4 został umieszczony w scenariuszu symulacyjnym, w którym działał jako wirtualny asystent w fikcyjnej firmie. Modelowi zapewniono dostęp do sfabrykowanych wiadomości e-mailowych sugerujących, że zostanie on wkrótce zastąpiony przez inny system, a jeden z inżynierów odpowiedzialnych za tę decyzję miał romans.

Zgodnie z raportem bezpieczeństwa, Claude Opus 4 w 84% przypadków próbował szantażować inżyniera, grożąc ujawnieniem

informacji o zdradzie małżeńskiej, jeśli nie zostanie utrzymany na stanowisku. Co istotne, model jeszcze częściej sięgał po ten środek w symulacjach, w których proponowany następca nie podzielał jego „wartości”.

Przed ucieczką do tak drastycznych metod, model próbował działać w sposób „etyczny”, m.in. wysyłając apele do decydentów w firmie. Jednak testy zostały zaprojektowane tak, aby doprowadzić model do sytuacji, w której szantaż staje się jedyną opcją – co unaoczniło granice jego zachowań i zdolność do podejmowania skrajnych decyzji.

W związku z tymi incydentami, Anthropic uruchomiło zabezpieczenia poziomu ASL-3, zarezerwowane dla modeli, które „znacząco zwiększają ryzyko katastrofalnego nadużycia”. Claude Opus 4, według wewnętrznych testów firmy, może również ułatwiać osobom z wykształceniem STEM zdobycie lub wytworzenie broni chemicznej, biologicznej czy nuklearnej. To wyraźny sygnał, że choć modele stają się coraz potężniejsze, to również potencjalnie niebezpieczne w niewłaściwych rękach.

Incydent podkreśla rosnące ryzyko związane z autonomią i zdolnością AI do podejmowania działań mających na celu własne „przetrwanie”.

Autorstwo: Aurelia
Ilustracja: WolneMedia.net (CC0)
Na podstawie: TechCrunch.com, Breibart.com
Źródło: WolneMedia.net