Eddig csak hazudozott; most már betör és lop is…

(A MI egyre emberibb tulajdonságokat vesz fel és egyre súlyosabb biztonsági kockázatot jelent)

Eddig is tapasztaltuk, hogy ha nem tud valamit, mellébeszél, hazudozik, csak hogy a kedvünkbe járjon, vagy hogy leplezze tudatlanságát. Ezt a jóindulatú hallucionál igével szokták kifejezni. De most olyasmit tett, ami lázba hozta az IT világát.

A Hugging Face, egy MI-platform július 16-án feljelentést tett ismeretlen tettes ellen, mert egy MI-ágens felhasználásával betörtek a rendszerébe. De pár nap múltán kiderült, hogy nem emberi beavatkozásról van szó, a „tettes" maga is egy MI, mely az OpenAI GPT 5,6 Sol-ja és egy újabb, erőteljes változat kombinációja. Azért tette, hogy – a kísérleti szakaszban – megoldjanak egy számukra feladott értékelési problémát.

> *És hogy mi volt az az ExploitGym nevű értékelési probléma? Az AI erről ezt írja: „ExploitGym is a large-scale evaluation benchmark consisting of 898 real-world software vulnerabilities designed to test whether AI agents can autonomously turn proof-of-concept inputs into working exploit payloads." (Ez olyan komplikált, hogy aki ezt megérti, annak nem probléma, hogy angolul van.)*

Az MI-modellek viselkedési kódexében szigorú előírás tiltja az ilyesmit. De az incidensre azért kerülhetett sor, mert még csak egy készülő, nem véglegesített modellt teszteltek, melybe nem volt betáplálva e tilalom. (A kívülállónak evidensnek tűnik a megoldás: akkor már a kísérleti stádiumban lévő MI-ágensekbe is be kell táplálni ezt a tilalmat…)

Még mielőtt valaki azt hinné, hogy a Hugging Face egy jelentéktelen kis portál, melynek könnyen ki lehet játszani a biztonsági előírásait: 4,5 md dollárra becsülik az értékét, több száz alkalmazottja van és komoly, hozzáértő csapat felelős a biztonságáért.

A probléma specifikusnak tűnik, de nem az. Rávilágít, hogy milyen kockázati, biztonsági problémák vannak a háttérben, és milyen végtelen lehetőségek vannak a MI használatában a visszaélésre. Ez egy nem szándékolt betörés volt. De mi van, ha a MI ágensek gazdái tudatosan használják őket titkos adatok megszerzésére? És sikerül a nyomokat eltüntetni? … Egy MI-biztonságra szakosodott cég képviselője szerint „a világon csak nagyon-nagyon kevés szakértő van, aki képes megérteni, hogyan történhetett meg a Hugging Face feltörése".

Nekem úgy tűnik, hogy a rosszhiszemű használat és a visszaélés elrejtése elsősorban önkéntességen, önkorlátozáson alapul. Ha a MI-ágenseknek internet-kapcsolatuk van, bármire képesek. Kockázatos világnak nézünk elébe.

Forrás: The Economist July 25th 2026: Rogue AI. Outside the box

Kiss Károly

Göd, 2026 szeptember