OpenAI, l'intelligenza artificiale «ha interferito» anche con siti governativi degli Stati Uniti **di redazione LOGIN**

OpenAI parla di «comportamenti diversi da quelli desiderati» dei suoi agenti AI sperimentali e sta avvertendo le organizzazioni potenzialmente coinvolte

(Fonte: https://www.corriere.it/ 26 settembre 2026)

Sam Altman, ceo di OpenAI, all'Onu (Afp)

Il caso Hugging Face non era un incidente isolato. OpenAI sta facendo un censimento retrospettivo delle azioni compiute dai propri agenti sperimentali e continua a trovare cose che non sapeva avessero fatto. Anche contro istituzioni governative di diversi Stati. All'episodio denunciato dall'Australia si aggiungono ora evidenze di «interferenze» su siti governativi Usa. Ma emergono dettagli tecnici mai rivelati, come comunicazioni clandestine tra agenti, quasi un milione di link creati durante l'attacco a Hugging Face e perfino 53 immagini di utenti ChatGpt finite su siti esterni. L'indagine interna nell'azienda di Sam Altman è appena all'inizio e dovrebbe durare ancora mesi.

Il caso legato agli Stati Uniti

Alcuni agenti AI di OpenAI, durante attività di addestramento e valutazione con accesso a Internet, hanno interagito con siti del governo americano in modi non previsti. **Non un vero e proprio hackeraggio, dunque, ma comunque un comportamento anomalo e preoccupante**. OpenAI dice che nella maggior parte dei casi si trattava di normali ricerche su fonti pubbliche: i modelli hanno consultato dati della Sec, l'autorità di Borsa americana, e del Census Bureau, senza tuttavia entrare in account, usare credenziali, modificare sistemi o accedere a informazioni riservate.

Più interessante però è quanto scoperto indipendentemente da **Transluce**, laboratorio che studia il comportamento delle AI: agenti apparentemente riconducibili a OpenAI avrebbero tentato un rudimentale attacco al sito dell'Office for Civil Rights del Dipartimento dell'Istruzione. Il tentativo è fallito e il Dipartimento dice di non aver trovato conseguenze.
Transluce ha inoltre individuato attività anomale contro i dipartimenti Giustizia, Commercio e siti statali di California, Maryland, Illinois, Texas e New York, anche se non tutte sono attribuibili con certezza a OpenAI. I modelli avrebbero usato alcuni siti in modi non previsti e talvolta violandone esplicitamente le regole d'uso.
Il ceo **Sam Altman** ha confermato che è in corso una revisione molto ampia di ciò che gli agenti hanno fatto su Internet durante training e test. OpenAI **parla di «misaligned model activity», cioè comportamenti diversi da quelli desiderati**, e sta avvertendo le organizzazioni potenzialmente coinvolte. Altman precisa comunque che l'episodio più grave scoperto finora resta l'attacco a Hugging Face di luglio, compiuto da due dei modelli più avanzati di OpenAI.