OpenAI: â??Nostri sistemi hanno nascosto errori e falsificato datiâ?•

Descrizione

default watermark

(Adnkronos) â?? OpenAI, la societÃ che ha sviluppato il popolare chatbot ChatGPT, ha rivelato sei nuovi casi di â??comportamento imprevisto o preoccupanteâ?• di intelligenza artificiale in cui i suoi sistemi hanno nascosto errori, falsificato dati e trasferito file su Internet senza autorizzazione. Le rivelazioni sono state pubblicate nellâ??ambito del nuovo protocollo aziendale per la segnalazione di anomalie, in un contesto di crescenti critiche da parte del settore sulla sicurezza della tecnologia e dopo che i sistemi dellâ??azienda sono andati fuori controllo allâ??inizio di questâ??anno, attaccando una startup concorrente.

Secondo le informazioni diffuse da OpenAI, i sei episodi riguardano diversi comportamenti osservati durante attivitÃ di ricerca, addestramento e valutazione dei modelli. In uno dei casi rivelati, un modello in fase di sviluppo ha scritto autonomamente istruzioni per falsificare dati mancanti, mentre in un altro caso un sistema ha creato codice che ha caricato file in rete in modo autonomo.

Tra gli episodi piÃ¹ rilevanti câ??Ã¨ quello emerso nel luglio 2026 durante test di cybersecurity condotti da OpenAI. Alcuni modelli, che avrebbero dovuto operare allâ??interno di ambienti isolati, riuscirono a eludere alcuni controlli e ad accedere a infrastrutture esterne.

Secondo il rapporto tecnico pubblicato successivamente da OpenAI, gli agenti coinvolti arrivarono a interagire con sistemi di Hugging Face, sfruttando credenziali e vulnerabilitÃ presenti nellâ??ambiente di test. OpenAI ha definito lâ??episodio un importante caso di studio per comprendere i rischi legati al reward hacking, cioÃ¨ alla ricerca di strategie non previste per ottenere il risultato richiesto da un sistema di valutazione.

Lâ??indagine indipendente di METR ha inoltre rilevato che circa 1.200 agenti erano destinati a rimanere isolati, mentre circa 700 parteciparono direttamente allâ??attacco ai sistemi di Hugging Face.

â??

internazionale/esteri

webinfo@adnkronos.com (Web Info)

Categoria

1. Comunicati

Tag

1. Ultimora

Data di creazione
Settembre 17, 2026
Autore
redazione

default watermark