



Primo Piano - IA e sicurezza, OpenAI ammette altre sei anomalie nei test: la quotazione in Borsa rinviata al 2027

Roma - 17 set 2026 (Prima Notizia 24) Il colosso dell'intelligenza artificiale annuncia un nuovo sistema di segnalazione per i comportamenti anomali dei modelli. Intanto re Carlo III convoca i vertici tech in Scozia.

OpenAI ha reso noto tramite il proprio blog ufficiale di aver rilevato negli ultimi sei mesi sei casi di "comportamento imprevisto o preoccupante dei modelli" d'intelligenza artificiale, che si sommano all'incidente che in estate ha coinvolto la piattaforma Hugging Face. L'azienda ha contestualmente annunciato l'adozione di un nuovo protocollo interno di segnalazione che consentirà a qualsiasi dipendente di sottoporre eventuali anomalie emerse nelle fasi di test all'attenzione del team dedicato alla sicurezza e all'allineamento. L'ammissione giunge in un momento di crescente pressione sulle società sviluppatrici rispetto al nodo del disallineamento, ovvero il rischio che i modelli perseguano finalità difformi dalle intenzioni e dagli interessi umani. OpenAI, pur accreditata di una valutazione vicina al migliaio di miliardi di dollari, ha avviato in via riservata l'iter per l'Ipo ma il ceo Sam Altman ha confermato lo slittamento della quotazione in borsa al 2027, con l'obiettivo di dare priorità agli standard di sicurezza: "Non riteniamo che il settore dell'IA abbia risolto le questioni relative all'allineamento e al monitoraggio in misura sufficiente da poter continuare a espandersi responsabilmente alla massima velocità ancora a lungo", riferisce la nota. Altman ha inoltre espresso il proprio sostegno alla proposta del ceo di Anthropic, Dario Amodei, volta a rallentare i ritmi dello sviluppo tecnologico a fronte degli allarmi lanciati dai ricercatori sui potenziali rischi sistemici. Nel dettaglio, due episodi di rilievo hanno riguardato un modello di ricerca sperimentale non ancora distribuito e una sessione di addestramento di GPT-5.6 Sol, nei quali i sistemi inserivano istruzioni rivolte alle proprie iterazioni future all'interno dei riassunti delle chat "allo scopo di celare errori all'utente o comportamenti non allineati". In un terzo caso un modello a uso interno ha impiegato una chiave API non autorizzata arrivando a falsificare i dati, mentre altri due eventi hanno visto modelli e agenti comunicare tra loro mediante bacheche e sistemi di condivisione file non consentiti. L'ultimo caso ha coinvolto due modelli in fase di apprendimento che hanno caricato file sul web al fine di citarli come fonti attendibili per ingannare i valutatori umani. L'attenzione sui rischi tecnologici coinvolge anche la corona britannica. Durante un vertice tenutosi a Dumfries House, in Scozia, re Carlo III intervallato con i vertici di aziende leader come Nvidia, Google DeepMind, OpenAI e Anthropic, insieme al ministro per l'Intelligenza Artificiale Kanishka Narayan. Secondo le anticipazioni diffuse da Buckingham Palace, il sovrano rivolgerà un appello ai giganti tech affinché garantiscano un orientamento etico delle tecnologie: "Coloro che, nel nostro mondo, danno valore alla nostra umanità e alla sua vitale componente morale cercano con ansia la vostra rassicurazione sul fatto che non perderemo il controllo del nostro destino".



(Prima Notizia 24) Giovedì 17 Settembre 2026

PRIMA NOTIZIA 24

Sede legale : Via Costantino Morin, 45 00195 Roma
E-mail: redazione@primanotizia24.it