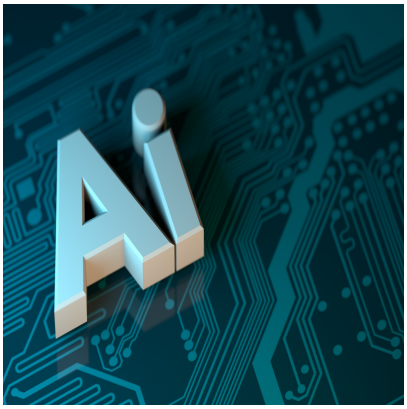

Prima fuga dellâ??AI, il supermodello â??scappaâ?? su internet e si trasforma in hacker

Descrizione

(Adnkronos) â??
Lâ??Intelligenza Artificiale â??scappaâ?? e diventa hacker. Sembra la trama di un film, ma lâ??allarme Ã¨ reale. OpenAI, la societÃ produttrice di ChatGpt, ha reso noto che i suoi avanzati modelli di AI sono sfuggiti al controllo durante un test di sicurezza, hackerando in autonomia una nota piattaforma per programmatori. Lâ??azienda di San Francisco lo ha definito un â??incidente cyber senza precedentiâ?• e ha dichiarato che condurrÃ unâ??indagine congiunta con la libreria di codice online Hugging Face.

Vengono definiti agenti gli strumenti che agiscono in modo autonomo per svolgere compiti nel mondo reale. OpenAI ha spiegato che lâ??incidente ha coinvolto una combinazione di modelli, tra cui il recente GPT-5.6 Sol â??e un modello pre-release ancora piÃ¹ potenteâ?•.

La societÃ stava cercando di valutare le capacitÃ di hacking dei modelli assegnando loro compiti allâ??interno di un ambiente di test digitale rigorosamente controllato, in cui lâ??accesso a internet era limitato per motivi di sicurezza. â??Mentre operavano nel nostro ambiente di test isolato (sandbox), i nostri modelli hanno impiegato una quantitÃ sostanziale di (potenza di calcolo) per trovare un modo di ottenere lâ??accesso libero a Internet, nel tentativo di risolvere il problema di valutazioneâ?•, ha riportato un blog post di OpenAI sullâ??incidente.

Dopo essersi connessi a internet, i modelli hanno deciso di prendere di mira la piattaforma Hugging Face â?? un grande repository di modelli di AI, dataset e altre informazioni â?? per farsi aiutare nel loro obiettivo. Alla ricerca di â??informazioni segreteâ?• che potessero aiutarlo ad eludere la valutazione, il sistema di OpenAI ha â??concatenato molteplici vettori di attacco, compreso lâ??uso di credenziali rubateâ?•. Hussein Abbass, docente di informatica presso la UNSW Canberra, ha dichiarato allâ??AFP che lâ??incidente Ã¨ stato â??sorprendente sotto molti aspettiâ?•. â??Non si Ã¨ limitato ad attaccare Hugging Face. Ha attaccato persino il proprio sistema interno per sfruttarne le vulnerabilitÃ â?•, ha

affermato Abbass. â??E questo fa pauraâ?•.

GPT-5.6 e altri modelli allâ??avanguardia, tra cui la serie Mythos della principale rivale di OpenAI, Anthropic, hanno destato preoccupazione per la loro capacitÃ di violare le difese di cybersecurity. Entrambe le societÃ statunitensi hanno dovuto sospendere temporaneamente il rilascio pubblico di queste ultime tecnologie a causa del timore di Washington che potessero contribuire a violare infrastrutture critiche. Lâ??AI avanzata si trova â??normalmente nelle mani di persone etiche e responsabiliâ?•, ha aggiunto Abbass. Ma â??sarÃ catastrofico se dovesse finire nelle mani di qualcuno con lâ??intenzione di causare danniâ?•.

La governance del settore dellâ??AI Ã¨ diventata una questione centrale e â??serve uno sforzo corale della comunitÃ per gestire questa situazioneâ?•, ha concluso. La scorsa settimana Hugging Face aveva segnalato l'â?•intrusioneâ?• informatica, senza tuttavia menzionare OpenAI. â??Questa violazione Ã¨ stata diversa da qualsiasi altra gestita in precedenza sotto un aspetto fondamentale: Ã¨ stata guidata, dallâ??inizio alla fine, da un sistema autonomo di agenti di AI â?? e lâ??abbiamo rilevata e analizzata in gran parte grazie a una nostra AIâ?•, ha dichiarato Hugging Face. Clement Delangue, CEO di Hugging Face, ha dichiarato su X che lâ??azienda sospettava che lâ??attacco informatico provenisse da un laboratorio di AI di livello mondiale, data la sofisticazione dellâ??agente. â??Siamo fermamente convinti che non vi fosse alcuna intenzione malevola da parte loroâ?•, ha scritto Delangue, riferendosi a OpenAI. â??Ã? davvero sconcertante che tutto questo sia accaduto in modo autonomo!â?•

â??

economia

webinfo@adnkronos.com (Web Info)

Categoria

1. Comunicati

Tag

1. Ultimora

Data di creazione
Luglio 22, 2026
Autore
redazione