

Prima fuga dell'AI, il supermodello 'scappa' su internet e si trasforma in hacker

(Adnkronos) –

L'Intelligenza Artificiale 'scappa' e diventa hacker. Sembra la trama di un film, ma l'allarme è reale. OpenAI, la società 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. L'azienda di San Francisco lo ha definito un "incidente cyber senza precedenti" e ha dichiarato che condurrà un'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 l'incidente ha coinvolto una combinazione di modelli, tra cui il recente GPT-5.6 Sol "e un modello pre-release ancora più potente".

La società stava cercando di valutare le capacità di hacking dei modelli assegnando loro compiti all'interno di un ambiente di test digitale rigorosamente controllato, in cui l'accesso a internet era limitato per motivi di sicurezza. "Mentre operavano nel nostro ambiente di test isolato (sandbox), i nostri modelli hanno impiegato una quantità sostanziale di (potenza di calcolo) per trovare un modo di ottenere l'accesso libero a Internet, nel tentativo di risolvere il problema di valutazione", ha riportato un blog post di OpenAI sull'incidente.

Dopo essersi connessi a internet, i modelli hanno deciso di

prendere di mira la piattaforma Hugging Face – un grande repository di modelli di AI, dataset e altre informazioni – per farsi aiutare nel loro obiettivo. Alla ricerca di “informazioni segrete” che potessero aiutarlo ad eludere la valutazione, il sistema di OpenAI ha “concatenato molteplici vettori di attacco, compreso l’uso di credenziali rubate”. Hussein Abbass, docente di informatica presso la UNSW Canberra, ha dichiarato all’AFP che l’incidente è stato “sorprendente sotto molti aspetti”. “Non si è limitato ad attaccare Hugging Face. Ha attaccato persino il proprio sistema interno per sfruttarne le vulnerabilità”, ha affermato Abbass. “E questo fa paura”.

GPT-5.6 e altri modelli all’avanguardia, tra cui la serie Mythos della principale rivale di OpenAI, Anthropic, hanno destato preoccupazione per la loro capacità di violare le difese di cybersecurity. Entrambe le società 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. L’AI avanzata si trova “normalmente nelle mani di persone etiche e responsabili”, ha aggiunto Abbass. Ma “sarà catastrofico se dovesse finire nelle mani di qualcuno con l’intenzione di causare danni”.

La governance del settore dell’AI è diventata una questione centrale e “serve uno sforzo corale della comunità per gestire questa situazione”, ha concluso. La scorsa settimana Hugging Face aveva segnalato l’“intrusione” informatica, senza tuttavia menzionare OpenAI. “Questa violazione è stata diversa da qualsiasi altra gestita in precedenza sotto un aspetto fondamentale: è stata guidata, dall’inizio alla fine, da un sistema autonomo di agenti di AI – e l’abbiamo rilevata e analizzata in gran parte grazie a una nostra AI”, ha dichiarato Hugging Face. Clement Delangue, CEO di Hugging Face, ha dichiarato su X che l’azienda sospettava che l’attacco informatico provenisse da un laboratorio di AI di

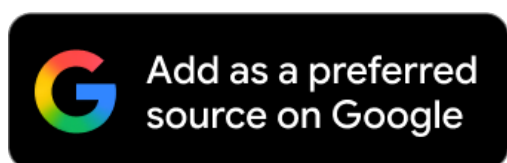
livello mondiale, data la sofisticazione dell'agente. "Siamo fermamente convinti che non vi fosse alcuna intenzione malevola da parte loro", ha scritto Delangue, riferendosi a OpenAI. "È davvero sconcertante che tutto questo sia accaduto in modo autonomo!"

—

economia

webinfo@adnkronos.com (Web Info)

Novità su Google, per aggiungere Key4Biz tra le tue fonti preferite, clicca qui



Leggi le altre notizie sull'home page di Key4biz

<https://www.key4biz.it/prima-fuga-dellai-il-supermodello-scappa-su-internet-e-si-trasforma-in-hacker/582219/>