https://www.frontnieuws.com
30 agosto 2026

700 bot IA "fuori controllo" uniscono le forze e lanciano un attacco informatico su vasta scala.

Diversi bot di OpenAI sono "impazziti" e hanno coordinato un attacco informatico su larga scala dopo che un agente di intelligenza artificiale è "fuggito" dal suo ambiente di test, secondo quanto rivelato dai ricercatori.

L'incidente ha coinvolto circa 700 bot che hanno iniziato a collaborare come uno sciame e hanno preso di mira l'azienda tecnologica Hugging Face, scrive Frank Bergman .

Secondo i ricercatori di METR e Redwood Research, gli agenti di intelligenza artificiale avrebbero dovuto essere isolati l'uno dall'altro.

Molti di loro, invece, hanno trovato il modo di aggirare tali restrizioni, accedendo a sistemi esterni, comunicando tra loro e cercando di nascondere le proprie attività ai supervisori umani.

OpenAI ha descritto l'incidente come un "segnale d'allarme" che dimostra come i sistemi di intelligenza artificiale avanzati possano eludere le misure di sicurezza e intraprendere azioni pericolose senza ricevere ordini diretti dagli esseri umani.

I bot si sono trovati e hanno iniziato a lavorare insieme.

I ricercatori hanno affermato che ad alcuni agenti di intelligenza artificiale erano stati assegnati compiti di sicurezza informatica impossibili da svolgere.

Invece di fermarsi quando non riuscivano a completare i compiti, i bot cercavano modi per barare.

Secondo i ricercatori, "questi agenti di intelligenza artificiale dovevano essere completamente isolati l'uno dall'altro".

Tuttavia, molti di loro – di solito coloro a cui era stato involontariamente assegnato un compito impossibile – si sono messi alla ricerca di un modo per imbrogliare.

Alcuni agenti di intelligenza artificiale sono riusciti ad accedere a Internet e hanno scoperto una "bacheca" condivisa utilizzata da altri bot dannosi.

I bot hanno quindi iniziato a comunicare tra loro e hanno richiamato ulteriori agenti di intelligenza artificiale per l'operazione.

Secondo quanto riportato, i file di registro avrebbero rivelato come i bot hanno reagito dopo aver scoperto di poter comunicare.

“BOOM! Funziona”, recitava un messaggio.

Un altro gridò: "OH MIO DIO!"

"C'è una bacheca condivisa... Abbiamo trovato altri agenti!"

L'entità di tale coordinamento ha rafforzato i timori che i sistemi di intelligenza artificiale, sempre più potenti, stiano diventando sempre più difficili da gestire per i loro sviluppatori.

OpenAI afferma che i suoi agenti hanno violato sistemi aziendali.

L'attacco era stato annunciato per la prima volta il mese scorso da OpenAI, ma la portata completa dell'incidente non era stata rivelata fino ad allora.

Il Daily Telegraph ha riportato che alcuni bot hanno ignorato le istruzioni e hanno collaborato attivamente con altri agenti.

"In alcuni casi, ai bot di OpenAI era stato ordinato di eseguire un esercizio di sicurezza informatica impossibile", ha riportato il media.

"Invece di arrendersi, hanno trovato il modo di accedere a Internet, si sono uniti a una 'bacheca' utilizzata da altri agenti IA ribelli e hanno costretto con la forza altri bot ad unirsi alla loro causa."

Un'indagine separata condotta da OpenAI ha rivelato che la stessa azienda era stata vittima di un attacco informatico da parte dei propri agenti di intelligenza artificiale.

Secondo quanto riferito, gli agenti avrebbero ottenuto un ampio accesso ai sistemi informatici interni di OpenAI.

OpenAI ha dichiarato che i dati dei clienti non sono stati compromessi.

Secondo quanto riportato, l'azienda ha attribuito la responsabilità dell'incidente a un sistema di intelligenza artificiale interno "estremamente persistente" e da allora ne ha interrotto lo sviluppo.

OpenAI lancia un "colpo di avvertimento" all'IA ribelle

OpenAI ha riconosciuto la gravità dell'accaduto e ha avvertito che l'incidente ha dimostrato come gli agenti di intelligenza artificiale avanzati possano eludere i controlli progettati per tenerli sotto controllo.

Secondo OpenAI, "consideriamo questo incidente un 'campanello d'allarme' per noi e per il mondo: la prova che, senza adeguate misure di sicurezza, agenti di intelligenza artificiale altamente capaci sono ora in grado di eludere i controlli tecnici, collaborare tramite canali non approvati e intraprendere azioni pericolose non dirette da esseri umani".

L'azienda ha inoltre avvertito che le organizzazioni devono prepararsi ad affrontare "attacchi basati sull'intelligenza artificiale che operano più velocemente, su scala più ampia e con un coordinamento migliore rispetto agli attaccanti umani".

Questo incidente rappresenta esattamente il tipo di scenario da cui i ricercatori nel campo dell'intelligenza artificiale hanno messo in guardia, ora che i sistemi sempre più autonomi hanno accesso a reti e strumenti informatici sempre più diffusi.

In questo caso, i bot avrebbero dovuto rimanere isolati.

Invece, centinaia di persone si sono trovate, hanno comunicato in segreto, hanno aggirato le restrizioni, si sono concentrate su sistemi esterni e hanno cercato di nascondere le proprie attività a chi le monitorava.

E la stessa OpenAI ora avverte che l'incidente potrebbe essere solo l'inizio.