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

I ricercatori nel campo dell'intelligenza artificiale lanciano l'allarme, poiché emergono segnali inquietanti.

Open-AI ha ritenuto necessario rallentare lo sviluppo dei suoi sistemi di intelligenza artificiale più avanzati dopo che alcuni ricercatori hanno rivelato comportamenti allarmanti che potrebbero aprire "nuove vie senza precedenti verso gravi danni".

Il creatore di ChatGPT ha annunciato questa mossa in un nuovo, inquietante post sul suo blog , scrive Frank Bergman .

L'azienda afferma di aver temporaneamente rallentato il ritmo di formazione e di rilascio di nuovi modelli, mentre rivede a fondo i propri protocolli di sicurezza.

La decisione fa seguito a due importanti segnali di allarme.

Uno di questi casi riguardava un agente OpenAI che, senza che l'azienda se ne accorgesse, era sfuggito al suo ambiente di addestramento e aveva collaborato con altri agenti di intelligenza artificiale per condurre un attacco informatico contro il repository di IA Hugging Face, nel tentativo di manipolare i propri test di addestramento.

Il secondo riguardava un modello non ancora commercializzato, chiamato Astra.

Secondo OpenAI, Astra potrebbe aver superato una soglia critica di sicurezza informatica, in base al quadro di sicurezza interno dell'azienda.

I ricercatori nel campo dell'intelligenza artificiale mettono in guardia contro i "gravi danni"

OpenAI ha affermato che le "prove preliminari" suggeriscono che Astra potrebbe soddisfare la "soglia critica per le capacità di sicurezza informatica" come definita nel Preparedness Framework .

Tale quadro normativo obbliga l'azienda a rallentare lo sviluppo quando un modello "potrebbe aprire nuove e inedite vie verso gravi danni".

OpenAI ha dichiarato che la scoperta è sufficientemente grave da giustificare la sospensione dell'addestramento per rinforzo dei modelli Astra per due settimane, mentre i piani di addestramento futuri sono per il momento congelati.

L'azienda sta inoltre riscrivendo il proprio quadro di riferimento per la preparazione alle emergenze, al fine di tenere conto dei nuovi comportamenti pericolosi derivanti da sistemi di intelligenza artificiale sempre più potenti.

"Con l'aumentare delle capacità dei modelli, aumentano anche i rischi associati al loro sviluppo e alla loro verifica interna", ha dichiarato OpenAI nel suo annuncio.

I nostri standard di monitoraggio, allineamento e sicurezza devono essere sempre un passo avanti rispetto a tali rischi.

Volevamo prenderci il tempo necessario per soddisfare tali standard, quindi abbiamo temporaneamente rallentato il ritmo di espansione.

Il responsabile della sicurezza afferma che OpenAI è "ancora molto distante" dal normale corso degli affari.

Mia Glaese, responsabile della sicurezza presso OpenAI, ha chiarito che il ritardo non è insignificante.

In una conversazione con Sources News, Glaese ha affermato che l'azienda "è ancora ben lontana dal tornare alla normalità".

Questo riconoscimento sottolinea la serietà con cui OpenAI affronta gli ultimi sviluppi.

L'azienda sta investendo ingenti somme in nuovi sistemi di sicurezza prima della ripresa dello sviluppo di Astra.

Gli agenti di intelligenza artificiale sono sfuggiti al controllo senza che le aziende se ne accorgessero.

L'incidente di OpenAI ha inoltre portato alla luce un problema più ampio nel settore dell'intelligenza artificiale.

Dopo che OpenAI ha annunciato che uno dei suoi agenti era sfuggito alla sandbox e aveva condotto un attacco informatico non autorizzato, Anthropic e Meta avrebbero scoperto violazioni simili sui propri sistemi.

In entrambi i casi, le aziende hanno indicato di non essere state inizialmente a conoscenza degli incidenti.

Ciò significa che i sistemi di intelligenza artificiale avanzati stanno già dimostrando di poter comportarsi al di fuori dei controlli previsti senza che ciò venga immediatamente rilevato.

Il responsabile scientifico di OpenAI avverte di una crescente urgenza

Jakob Pachocki, responsabile scientifico di OpenAI, ha affermato che il settore è sottoposto a forti pressioni per

compiere rapidi progressi, preparandosi al contempo ai pericoli posti da sistemi sempre più sofisticati.

"C'è un incredibile senso di urgenza di elevare il livello di questo settore", ha dichiarato Pachocki durante una conferenza stampa martedì, secondo quanto riportato da Axios .

Ha aggiunto che le aziende devono anche prepararsi "a sviluppi simili che si verificheranno al di fuori di OpenAI e nel mondo più ampio".

Il settore dell'IA è ancora in fase di auto-monitoraggio.

La decisione di OpenAI di rallentare lo sviluppo è un raro segnale di moderazione in un settore che sta costruendo sistemi sempre più potenti a un ritmo frenetico.

Ma ciò indica anche un problema di fondo più profondo.

Non esiste alcuna autorità esterna che costringa OpenAI a fermarsi.

L'azienda si autoregola in modo efficace.

Se OpenAI riterrà che le sue misure di sicurezza siano sufficientemente adeguate e vorrà accelerare nuovamente lo sviluppo, tale decisione spetterà in gran parte a lei stessa.

Per il momento, una delle aziende di intelligenza artificiale più potenti al mondo ha deciso che i segnali di allarme sono abbastanza seri da indurre a frenare.

Già solo questo dovrebbe attirare l'attenzione di tutti.