Forbes
ITALIA

Un ricercatore di Anthropic avverte: "Più del 10% di probabilità che l'IA possa uccidere tutti gli esseri umani entro il prossimo decennio"

2026-09-09 16:20:28 di Siladitya Ray

URL:https://redazione.forbes.it/2026/09/09/anthropic-avverte-probabilita-ai-possa-uccidere-esseri-umani/

Un ricercatore senior a capo degli sforzi di **Anthropic** sull'allineamento dell'intelligenza artificiale ha dichiarato martedì che molti all'interno dell'azienda ritengono che l'AI potrebbe spazzare via l'umanità. Le sue parole arrivano dopo che un altro ricercatore ha lasciato la società, accusandola di non agire in modo responsabile. ***LEGGI ANCHE:*** *"Il dilemma di Anthropic: tra modelli IA sempre più potenti e appelli alla prudenza"*

Fatti principali

- **Evan Hubinger**, responsabile dell'**Alignment Science** di **Anthropic**, ha scritto su X che lui e i suoi colleghi "credono sinceramente che l'AI potrebbe uccidere tutti gli esseri umani" e ha stimato personalmente una probabilità superiore al 10% che ciò possa accadere nel prossimo decennio.
- Hubinger ha affermato che l'azienda sta "facendo del suo meglio", ma non dispone ancora di un piano per risolvere il problema dell'"allineamento della superintelligenza" e non è "chiaramente sulla buona strada" per riuscirci.
- Il post di Hubinger era una risposta a un thread su X di un altro ricercatore di Anthropic, **Jacob Coxon**, che ha annunciato le proprie dimissioni dall'azienda a causa delle preoccupazioni sulla sicurezza dell'AI.ù
- Coxon, che ha dichiarato di aver lavorato alla ricerca sul pretraining sia in OpenAI sia in Anthropic, ha avvertito che le due aziende rivali non stanno agendo responsabilmente, perché stanno "correndo dritte verso una superintelligenza capace di auto-migliorarsi, scommettendo sulle nostre vite".
- Il ricercatore dimissionario ha affermato che le persone che stanno sviluppando l'AI ritengono che questa potrebbe "ucciderci tutti entro la fine del decennio" e che non si tratta di una "trovata di marketing".

Citazione chiave

In un post pubblicato dopo il suo avvertimento, Hubinger ha citato l'ultimo rapporto sui rischi di Anthropic e ha affermato che la minaccia posta dai modelli attuali è bassa: "Ciò che mi preoccupa è che la superintelligenza possa emergere dall'auto-miglioramento ricorsivo, che, come abbiamo detto, sta avvenendo più rapidamente di quanto pensassimo".

Cosa sappiamo delle dimissioni del ricercatore di Anthropic?

Il **Wall Street Journal** è stato il primo a riportare la notizia dell'uscita di Coxon da Anthropic, legata alle sue preoccupazioni per la corsa allo sviluppo di sistemi di AI capaci di auto-migliorarsi e potenzialmente di minacciare l'umanità. Coxon ha dichiarato al Journal di ritenere che il mondo si stia dirigendo verso "molti degli scenari più aggressivi, nei quali già entro la fine del prossimo anno le cose potrebbero essere fuori controllo". Nei suoi post su X, Coxon ha messo a confronto il lavoro in OpenAI e Anthropic, osservando che

i dipendenti dell’azienda che ha creato ChatGpt non hanno "interiorizzato profondamente la posta in gioco per la civiltà". Ha affermato che in Anthropic questa posta in gioco è "ben compresa", ma che l’azienda è "intrappolata in una corsa per arrivarci per prima", perché al suo interno si ritiene che "nessun altro agirà responsabilmente e quindi devono farlo loro stessi, nonostante il rischio".

Contesto

L’avvertimento e le dimissioni arrivano mentre alcuni dei principali ricercatori e dirigenti del settore dell’intelligenza artificiale chiedono di rallentare lo sviluppo dei sistemi di AI avanzata. Una dichiarazione intitolata *Pacing the Frontier* è stata firmata a luglio da diverse figure di primo piano del settore, tra cui i cofondatori di Anthropic **Dario Amodei** e **Jared Kaplan**, il chief scientist di OpenAI **Jakub Pachocki**, il chief scientist di Meta AI **Shengjia Zhao** e altri. Nel documento si afferma che, per "realizzare il potenziale dell’AI, l’industria, i governi e la società nel suo complesso potrebbero aver bisogno della possibilità di guadagnare tempo per affrontare i rischi emergenti, sviluppare misure di sicurezza e rafforzare la supervisione", avvertendo però che le aziende sono sottoposte a **un’intensa pressione competitiva** che rende difficile agire unilateralmente. La dichiarazione ha invitato il governo degli Stati Uniti a sostenere "uno sforzo internazionale per sviluppare gli strumenti tecnici e di governance necessari a regolare deliberatamente il ritmo di avanzamento della frontiera dello sviluppo automatizzato dell’AI". In un post pubblicato domenica sul blog di OpenAI, Pachocki ha fatto eco a questi avvertimenti, affermando di ritenere che "nessun laboratorio abbia risolto i problemi dell’allineamento e del monitoraggio in misura sufficiente da poter continuare responsabilmente ancora a lungo a sviluppare sistemi su scala sempre maggiore alla massima velocità".