Fuga dai Chatbot – Chi crea l'AI sa che potrebbe sterminare l'umanità entro la fine del decennio

Pubblicato il 09/09/2026

Un nuovo e preoccupante allarme sul futuro dell'intelligenza artificiale arriva direttamente dall'interno di Anthropic, la società che ha sviluppato il chatbot Claude. Dopo aver rassegnato le dimissioni, tre ex ricercatori dell'azienda hanno deciso di denunciare pubblicamente i rischi legati allo sviluppo di questa tecnologia, affermando che gli stessi addetti ai lavori temono conseguenze catastrofiche per l'umanità.

Le dimissioni e le denunce sui social

L'ex dipendente Jacob Coxon ha dichiarato apertamente sul social X che chi sviluppa l'intelligenza artificiale crede sinceramente nella possibilità che questa possa uccidere tutti gli esseri umani entro la fine del decennio. Secondo Coxon, non si tratta di una trovata di marketing: sebbene molti dirigenti e ricercatori di alto livello tendano a usare un linguaggio più

pacato con la stampa, privatamente esprimono la medesima paura, considerando l'AI un pericolo senza eguali rispetto a qualsiasi altra attività umana.

A sostenere questa tesi sono intervenuti anche altri due ex colleghi di Anthropic:

- **Evan Hubinger**, responsabile scientifico per l'allineamento (il processo volto a guidare i sistemi affinché i loro comportamenti siano conformi ai valori e alle esigenze umane), ha confermato le parole di Coxon stimando che la probabilità di uno sterminio dell'intera umanità supererà il 10% nel prossimo decennio. Hubinger ha inoltre sottolineato che, pur facendo Anthropic del suo meglio, non esiste ancora un piano risolutivo per il problema dell'allineamento della superintelligenza.
- **Samuel Marks**, responsabile della supervisione scalabile, ha aggiunto che all'interno della categoria la preoccupazione cresce in proporzione diretta con l'anzianità e il livello di esperienza del dipendente.

> *I resigned from Anthropic today. I spent the last three years doing pretraining research at both OpenAI and Anthropic. Neither company is acting responsibly. They are racing straight to self-improving superintelligence and gambling with our lives. More thoughts below.*
>
> *— Jacob Coxon (@hilbertspaess) September 9, 2026*

Il dilemma dello sviluppo e la corsa globale

Il dibattito sulla pericolosità dell'intelligenza artificiale si scontra tuttavia con una dinamica complessa: fermare o rallentare la ricerca significherebbe lasciare campo libero a concorrenti, sia occidentali che non, pronti a far progredire i propri modelli senza freni. Nel frattempo, i progressi tecnologici corrono veloci, come dimostrano i traguardi di OpenAI verso l'AGI (intelligenza artificiale generale) e i recenti episodi in cui gli agenti di AI hanno dimostrato di saper coordinarsi a qualsiasi costo, in particolare nell'ambito della sicurezza informatica.

Queste denunce da parte degli ex dipendenti di Anthropic si inseriscono in un contesto di crescenti pressioni internazionali, giungendo a pochi giorni di distanza dall'appello delle Nazioni Unite — a firma dell'alto commissario per i diritti umani Volker Türk — in cui si esortano gli Stati ad agire tempestivamente prima che l'intelligenza artificiale diventi un rischio esistenziale fuori

controllo.

Link notizia: https://www.corriere.it/tecnologia/26_settembre_09/l-allarme-del-ricercatore-di-anthropic-l-ai-potrebbe-sterminare-gli-esseri-umani-8f16ad30-6f3d-4969-9a6a-9d30bbacfxlk.shtml