L'allarme del ricercatore di Anthropic: «L'AI potrebbe sterminare gli esseri umani» di Roberto Cosentino

Dopo le dimissioni, l'ex dipendente della società che ha sviluppato Claude ha deciso di raccontare pubblicamente che «Le persone che sviluppano l'AI credono sinceramente che potrebbe ucciderci tutti entro la fine del decennio». La conferma da altri due suoi ex colleghi (Fonte: https://www.corriere.it/ 9 settembre 2026)

L'ennesimo allarme arriva da Anthropic, la società che ha sviluppato il chatbot Claude: «**Sì, l'AI potrebbe sterminare l'umanità**». Ciclicamente gli esperti di intelligenza artificiale avvertono di quella catastrofica eventualità secondo cui l'AI potrebbe diventare un grosso pericolo per la nostra stessa esistenza. Decenni di antologia cinematografica non sono bastate: tutto ad un tratto, Skynet potrebbe non essere più relegata alla fortunata serie cinematografica *Terminator*.
Il nuovo allarme viene scritto e pubblicato da **tre ricercatori** che lavoravano nella società dei fratelli Amodei. Dopo aver rassegnato le dimissioni, hanno deciso di esternare le loro preoccupazioni sul social X. Uno di questi, **Jacob Coxon**, ha affermato che «le persone che sviluppano l'AI **credono sinceramente che potrebbe ucciderci tutti entro la fine del decennio.** Non si tratta di una trovata di marketing. Anzi, molti dirigenti e ricercatori di alto livello cercheranno di usare un linguaggio più pacato con la stampa, ma sento le stesse persone esprimere privatamente questa paura. Nessun'altra attività umana rappresenta un pericolo di questo livello».
A confermare le preoccupazioni di Coxon anche **Evan Hubinger,** responsabile scientifico per l'allineamento sempre ad Athropic. Per «allineamento» si intende il processo per guidare i sistemi

AI affinché i loro comportamenti, obiettivi e output siano conformi ai valori, alle intenzioni e alle esigenze degli esseri umani. «Jacob ha ragione - ha risposto - crediamo davvero che l'AI potrebbe sterminare tutta l'umanità! Personalmente, penso che la percentuale supererà il 10% entro il prossimo decennio. **Credo che Anthropic stia facendo del suo meglio, ma non abbiamo ancora un piano per risolvere il problema dell'allineamento** per la superintelligenza e non siamo chiaramente sulla buona strada per farlo». Sulla stessa linea anche **Samuel Marks**, responsabile della supervisione scalabile di Anthropic, che aggiunge: «Gli sviluppatori di AI ritengono che la loro tecnologia potrebbe causare l'estinzione umana (o conseguenze altrettanto negative). Ciò potrebbe accadere nei prossimi anni. In generale, più alto è il livello di anzianità del dipendente, maggiore è la sua preoccupazione».

Come dicevamo, ciclicamente si torna al punto in cui chi è coinvolto nei vari sviluppi, lancia allarmi e preoccupazioni. Da quando l'AI è diventata di uso comune, sin dal lancio di ChatGpt, sono nati diversi progetti volti a monitorare e talvolta, a livello governativo, bloccare le capacità dei modelli. Ma perché semplicemente non rallentare o spegnere tutto?

La risposta a questa domanda è ovvia, ma non c'entrano solo i soldi. Gli allarmi, i dubbi e le perplessità, sono alimentate perlopiù dagli sviluppatori occidentali.

Fermarsi o rallentare, **lascerebbe via libera a rivali o concorrenti, occidentali e non,** a migliorare di volta in volta i propri modelli, senza che vi siano freni o limiti. La verità è che il miglioramento dei modelli procede a gonfie vele e proprio recentemente, **OpenAI ha affermato di aver raggiunto il livello di AGI** (intelligenza artificiale generale) ovvero punto di non ritorno in cui l'AI supererebbe le capacità umane. Ma il vero pericolo potrebbe emergere qualora l'AI **diventasse in grado di migliorarsi da sé.** Gli ultimi sviluppi che hanno coinvolto proprio gli agenti di AI hanno dimostrato di sapersi coordinare per raggiungere un obiettivo a qualsiasi costo, **in particolar modo nei casi di sicurezza informatica**. **L'incidente ad HuggingFace,** di cui i maggiori dettagli sono emersi solo negli ultimi giorni, ne sono un esempio.

Gli allarmi lanciati dagli ex dipendenti di Anthropic arrivano a pochi giorni di distanza dall'**appello delle Nazioni Unite** a firma dell'alto commissario per i diritti umani, Volker Türk, in cui chiedeva agli Stati di intervenire prima che l'AI vada fuori controllo e che diventi «un rischio esistenziale per l'umanità».