La tecnologia sfuggirà al controllo? E quali sono i rischi per la vita umana? Domande e risposte di Federico Cella

Cosa sono automiglioramento e disallineamento

(Fonte: https://www.corriere.it/ 13 settembre 2026)

Il 9 settembre il ricercatore Jacob Coxon, 27 anni, si è dimesso da Anthropic per sentirsi libero di denunciare pubblicamente quelli che a suo parere sono i pericoli derivanti da **uno sviluppo poco controllato** dei modelli più avanzati di intelligenza artificiale.

1 Che cosa ha scritto Jacob Coxon sui social?

L'informatico ha lavorato tre anni all'allenamento dei modelli di AI, prima a OpenAI quindi ad Anthropic. «Nessuna delle due aziende sta agendo in modo responsabile... stanno giocando d'azzardo con le nostre vite». A suo sostegno altri due ricercatori, tra cui Evan Hubiger, responsabile scientifico per l'allineamento — il «comportamento» di un modello — sempre ad Anthropic: «Crediamo davvero che l'AI potrebbe sterminare tutta l'umanità! Personalmente, penso che la percentuale che questo possa accadere supererà il 10% entro il prossimo decennio».

2 Che cosa si intende per un'AI capace di automigliorarsi?

S'intende un sistema in grado di contribuire alla progettazione di modelli più evoluti di sé stesso, accelerando progressivamente la ricerca. Il rischio temuto da Coxon è che si inneschi un circuito sempre più rapido.

3 Quali sono i motivi per cui l'AI potrebbe diventare pericolosa per l'umanità?
Bisogna innanzitutto distinguere tra un uso malevolo dell'AI (artificial intelligence), ossia un utilizzo criminale da parte di esseri umani, e invece la perdita di controllo dei modelli da parte dei loro sviluppatori. Focalizzandosi sul secondo punto, quello che tecnicamente potrebbe accadere è un «disallineamento» dei modelli, cioè l'andare oltre i principi e i valori impostati per raggiungere l'obiettivo prefissato.

4 Un esempio di «disallineamento»?
Eliminare chi inquina, cioè l'uomo, per risolvere la crisi climatica. Un altro rischio è l'«autotutela»: la macchina potrebbe decidere di azzerare gli ostacoli che si frappongono tra lei e la soluzione di un problema. La terza possibilità di deviazione è legata alla «delega estrema», per esempio affidare all'intelligenza artificiale la gestione di arsenali militari automatici.

5 Come hanno reagito le principali aziende che sviluppano AI?
La prima realtà a scendere in pista è stata OpenAI, con una richiesta ufficiale al Congresso americano di introdurre requisiti obbligatori per i modelli di AI più avanzati. Sam Altman, amministratore delegato della società, in un discorso interno avrebbe manifestato la volontà di concordare con le altre aziende il rallentamento del tasso di velocità della ricerca sui nuovi modelli. Pochi giorni dopo gli ha fatto eco Dario Amodei, ceo di Anthropic: «Dobbiamo rallentare il ritmo con cui miglioriamo le capacità dei modelli. I progressi sembreranno comunque rapidi e dovremo fare un uso saggio del tempo che guadagneremo».

6 La comunità tecnologica ha reagito compatta agli allarmi?
Inizialmente Elon Musk, a capo di xAI, aveva bollato i proclami dei ricercatori come «una montatura». L'azienda dell'imprenditore naturalizzato americano licenziò lo scorso giugno un ingegnere per avere segnalato problemi di sicurezza legati allo sviluppo di Grok. Ieri però ha fatto dietrofront: «Dario ha ragione». Altri commentatori hanno definito i proclami come vero marketing, perché le due aziende stanno per sbarcare in Borsa.

7 Questi allarmi sono i primi a essere denunciati pubblicamente?
No, nella storia molto recente dell'AI, i primi dubbi sono sorti quasi subito. A rappresentarli, uno dei cosiddetti «padrini» dell'intelligenza artificiale moderna, Geoffrey Hinton: il Nobel nel 2024 lasciò il lavoro in Google proprio per denunciare i pericoli derivanti sia da un utilizzo da parte di «attori maligni», sia dalla perdita di controllo. Già allora un documento firmato da oltre mille tra sviluppatori e imprenditori chiedeva una moratoria nello sviluppo dell'AI. Raccolte di firme (illustri) che sono proseguite anche negli scorsi luglio e agosto.

8 Esistono regole per contenere i rischi?

La Ue è intervenuta con l'AI Act, che impone obblighi specifici ai produttori dei modelli più avanzati considerati a «rischio sistemico» ma la legge non impedisce alle aziende di sviluppare sistemi sempre più potenti.