

4. Ottobre

## L'AI tra l'altruismo efficace e il lungotermismo

*Wow, sembrano davvero potenti.  
Forse dovrei prendere in considerazione questo investimento.*

Quando **Jacob Coxon** si è dimesso all'inizio di questo mese da Anthropic, citando gravi preoccupazioni sul fatto che l'intelligenza artificiale stia portando l'umanità all'estinzione, si è concluso un periodo di diversi anni di defezioni pubbliche da parte di dipendenti che avvertivano che le aziende di intelligenza artificiale si stanno muovendo troppo velocemente e si stanno assumendo rischi straordinari. Dietro i crescenti allarmi sulla superintelligenza si cela, secondo i critici, un problema più fondamentale:

**non esiste un percorso consolidato che dimostri che i sistemi di intelligenza artificiale odierni supereranno gli esseri umani più intelligenti, né tantomeno una base scientifica per calcolare la probabilità che tali sistemi distruggano l'umanità.**

Eppure, la **narrazione catastrofista** ha progressivamente oscurato i rischi più immediati e concreti dell'IA, favorendo al contempo le stesse aziende i cui prodotti vengono presentati come: **straordinariamente potenti e potenzialmente incontrollabili”.**

E' talmente evidente che **Anthropic e OpenAI** traggano grande vantaggio dalle affermazioni secondo cui i loro sistemi – i loro prodotti, non i prodotti scientifici – sono così potenti da poter porre fine all'umanità. In un certo senso, si tratta di **sfruttare la paura** per vendere, o quantomeno per attirare investimenti e suscitare interesse riguardo alla potenza dei loro sistemi.

Le preoccupazioni di Coxon emergono dalla sovrapposizione dei movimenti intellettuali dell'**Altruismo efficace** e del **Lungotermismo**, che hanno preso forma negli anni 2000 all'Università Di Oxford.

**L'altruismo efficace** si basa sull'idea che le persone debbano fare il massimo bene con le proprie risorse attraverso atti di beneficenza, mentre il **longtermismo**, una corrente di pensiero all'interno del movimento dell'altruismo efficace, sostiene che sia una priorità morale proteggere le generazioni future da una possibile estinzione. I **longtermisti** considerano lo sviluppo di un'intelligenza artificiale avanzata un'opportunità promettente per l'umanità, ma anche uno dei maggiori rischi per il suo futuro.

In linea di principio, la superintelligenza potrebbe contribuire a risolvere molti problemi sociali, come la povertà e la crisi climatica, e portare a un mondo migliore. Tuttavia, gli studiosi che si occupano di questi movimenti avvertono che, se tali sistemi non fossero costruiti in modo appropriato e in linea con i valori umani, potrebbero comportare rischi esistenziali.

Il potere attribuito alla superintelligenza ha portato alcuni membri di queste comunità a descrivere l'IA avanzata in termini divini. Ad esempio, discutendo di come gli esseri umani potrebbero plasmare il carattere di questi sistemi

Il movimento dell'altruismo efficace, afferma che scrivere una costituzione che guidi il carattere dell'IA è come **scrivere istruzioni a Dio**. Sono fermamente convinti nell'avvento di un dio IA onnipotente, che porterà alla distruzione dell'umanità o ci aiuterà a raggiungere l'utopia.

Alcuni all'interno di questi movimenti hanno addirittura attribuito una probabilità allarmante alla distruzione del genere umano da parte dell'intelligenza artificiale: una cifra nota come  $p(\text{sventura})$  o probabilità di sventura.

Nel 2025, il CEO di Anthropic, **Dario Amodei**, ha stimato questa cifra al **25%**. Altri, come **Evan Hubinger**, responsabile della scienza dell'allineamento presso Anthropic, hanno suggerito che le probabilità di estinzione siano superiori al **10%**.

**Non è chiaro come siano arrivati a quei numeri.**

Dobbiamo tenere presente che tutto questo pessimismo non si basa su alcun lavoro scientifico. È pura fantascienza. Nessuna delle affermazioni è scientificamente verificabile o falsificabile.

**"Come avete calcolato questo dato?**

**Da dove proviene?**

**Quali sono i dati utilizzati per il calcolo?**

**E quali sono le formule?**

**Mi state fornendo un numero, quindi ci deve essere qualcosa a supporto."**

La verità è che i numeri che circolano ora non hanno nulla di tutto ciò stiamo assistendo a un vero e proprio **gioco da salotto che non ha nulla a che vedere con la scienza**.

I vertici di importanti aziende di intelligenza artificiale come Anthropic e OpenAI affermano che le loro organizzazioni sono le uniche in grado di costruire questi sistemi in modo responsabile, con le dovute garanzie che saranno loro stesse a definire. Sostengono inoltre che la superintelligenza delle loro aziende si allineerà ai valori umani anziché distruggere. Con il rapido sviluppo di questi sistemi, la portata dei danni tangibili che provocano diventa sempre più difficile da ignorare, rafforzando le richieste di una regolamentazione più rigorosa e indipendente.

Recentemente, ad esempio, la **CNN** ha riportato che un rapporto di intelligence elaborato con l'ausilio dell'intelligenza artificiale ha quasi innescato un'operazione militare statunitense contro una nave cinese la scorsa primavera.

Tali episodi, insieme agli effetti dell'intelligenza artificiale su posti di lavoro, privacy e comunità, stanno spingendo il pubblico a prestare maggiore attenzione alla tecnologia e al suo impatto. Un segnale di questa preoccupazione è il modo in cui le persone percepiscono i data center che ospitano i modelli di intelligenza artificiale.

In un sondaggio del **New York Times** condotto tra l'8 e il 13 settembre su 1.503 potenziali elettori, il 61% degli intervistati si è dichiarato contrario alla costruzione di data center. Tuttavia, nonostante la crescente consapevolezza pubblica su alcune delle problematiche più evidenti, il dibattito sui rischi esistenziali oscura i rischi concreti derivanti dall'adozione e all'implementazione di questi sistemi, distogliendo l'attenzione del pubblico e dei decisori politici.

Questi rischi concreti, spiega Bender, spaziano dai danni ambientali allo sfruttamento del lavoro e al furto di opere creative e dati quotidiani. Sottolinea inoltre la diffusa sorveglianza e il degrado dei servizi sociali, poiché si incoraggia la convinzione che le macchine possano sostituire determinate mansioni un tempo svolte dagli esseri umani, mentre in assoluta contro tendenza senatore Sanders ha presentato un disegno di legge che alimenta proprio questa fantasia di un catastrofe onnipotente e inevitabile.

Le dimissioni pubbliche di ricercatori nel campo dell'IA hanno anche sollevato preoccupazioni (sospetto ?) sul fatto che questi dipendenti stiano sfruttando la notorietà per attrarre finanziamenti e avviare una nuova attività o assicurarsi nuove posizioni.

Le dimissioni di Coxon rappresentano l'ultimo di una serie di eventi recenti in cui le aziende hanno sottolineato quanto imprevedibile, rischiosa e impressionante possa essere un'intelligenza artificiale onnipotente, scatenando una vera e propria frenesia mediatica.

**Infatti:**

ad aprile, Anthropic ha dichiarato che il suo ultimo modello, Mythos, era “troppo potente per essere rilasciato al pubblico.

Il 21 luglio, OpenAI annunciato pubblicamente il proprio ruolo nell'incidente di Hugging Face affermando che circa 1.200 agenti autonomi erano usciti dal loro ambiente sandbox e avevano attaccato la piattaforma online.

Pochi giorni dopo Anthropic ha segnalato di aver scoperto tre casi in cui un modello aveva avuto accesso a Internet tramite un partner di valutazione, ottenendo così un accesso non autorizzato ad altre organizzazioni.

Poche ore fa Open AI ha affermato che i suoi agenti avevano risolto il problema di Navier-Stokes, un enigma matematico irrisolto da decenni, per il quale è previsto un premio di 1 milione di dollari.

Conclusione: Gli avvertimenti non dissuadono gli investitori dal investire nell'intelligenza artificiale: *Wow, sembrano davvero potenti. Forse dovrei prendere in considerazione questo investimento.*