Le intelligenze artificiali dovrebbero avere diritti?

Il mondo della ricerca sull'**intelligenza artificiale** è spesso bizzarro. Oggi nella Silicon Valley c'è un campo ancora piccolo ma in crescita, chiamato ***model welfare***, che sta cercando di **capire se i modelli AI sono coscienti e meritano di essere oggetto di considerazioni morali**, come **i diritti giurdici**. Nell'ultimo anno sono nate due organizzazioni finalizzate a esplorare questo tempo – Conscium e Eleos AI

Research – e nel 2024 Anthropic ha assunto il suo primo ricercatore specializzato in benessere delle AI.

All'inizio di settembre, la società fondata dai fratelli Amodei ha dichiarato di aver dotato il suo chatbo, Claude, della capacità di **terminare le "*interazioni dannose o abusive con gli utenti*" che potrebbero rivelarsi "*potenzialmente angoscianti*"**.

"*Rimaniamo molto incerti sul possibile status morale di Claude e di altri llm, ora o in futuro* – ha dichiarato Anthropic in un post pubblicato sul blog aziendale –. *Tuttavia, **prendiamo la questione sul serio** e, insieme al nostro programma di ricerca, stiamo lavorando per identificare e implementare interventi a basso costo per mitigare i **rischi per il benessere dei modelli***".

Anche se i timori per la salute dell'intelligenza artificiale potrebbero sembrare ridicoli, **l'idea non è nuova**. Più di mezzo secolo fa, il matematico e filosofo americano Hilary Putnam, per esempio, si chiedeva se i robot dovessero godere di diritti civili. "*Dato il ritmo sempre più accelerato dei cambiamenti tecnologici e sociali, è del tutto possibile che un giorno i robot esisteranno, e che sostengano: 'Siamo vivi, siamo coscienti!'*", scrisse Putnam in un articolo del 1964.

Oggi, a distanza di molti decenni, i **progressi dell'intelligenza artificiale hanno prodotto effetti ancora più strani** di quelli che il filosofo avrebbe mai potuto prevedere. Le persone si innamorano dei chatbot, ipotizzano che possano provare dolore e trattano la tecnologia come una divinità in grado di attraversare lo schermo. Per non parlare dei funerali ai modelli AI e delle feste organizzate per discutere su come potrebbe essere il mondo una volta che le macchine avranno ereditato la Terra.

In modo forse sorprendente, i **ricercatori che si occupano di *model welfare* sono tra quelli che si oppongono all'idea che**

le AI debbano essere considerate coscienti, almeno per il momento. Rosie Campbell e Robert Long, che collaborano alla guida di **Eleos AI**, un'organizzazione di ricerca no-profit dedicata al benessere dei modelli, mi hanno raccontato di ricevere molte email da persone che sembrano completamente convinte che le intelligenze artificiali siano già senzienti. I due hanno persino contribuito alla stesura di una guida per le persone preoccupate dalla possibilità di un'AI senziente.

"*Uno schema comune che notiamo in queste email sono le **persone che affermano che esista un complotto per eliminare le prove della coscienza** – spiega Campbell –. Penso che se noi, come società, reagiamo a questo fenomeno rendendo tabù anche solo prendere in considerazione la questione e chiudendo in un certo senso ogni dibattito a riguardo, stiamo essenzialmente facendo sì che quest complotto si avveri*".

Nessuna prova di intelligenza artificiale cosciente

La mia reazione iniziale quando ho scoperto per la prima volta dell'esistenza del campo è potrebbe simile alla vostra. Dal momento che il mondo è a malapena in grado di dare il giusto valore alla vita degli esseri umani reali e di altri esseri coscienti come gli animali, **attribuire una personalità a macchine probabilistiche potrebbe apparire davvero fuori luogo**. È un aspetto che Campbell dice di aver preso in considerazione. "*Visti i nostri precedenti storici di sottovalutazione dello status morale di diversi gruppi e vari animali, penso che dovremmo essere molto più umili e cercare di rispondere effettivamente alla domanda*".

In un paper, da Eleos AI sostiene la necessità di valutare la coscienza delle AI utilizzando un **approccio improntato al "*funzionalismo computazionale*"**, un'idea simile a quella sostenuta dallo stesso Putnam (che però la criticò in una fase successiva della sua carriera). La teoria suggerisce che la

menti umane possono essere considerate come specifici tipi di sistemi computazionali. Questa premessa permetterebbe di capire se altri sistemi, come un chabot, sono dotati indicatori che suggeriscono la presenza di una coscienza simile a quella di un essere umano.

Nel documento, Eleos AI afferma che "*una delle principali sfide per l'applicazione*" di questo approccio è rappresentata dal fatto "*che comporta una serie di **rilevanti decisioni discrezionali**, sia nella formulazione degli indicatori che nella valutazione della loro presenza o assenza nei sistemi di intelligenza artificiale*".

Il *model welfare* è ovviamente ancora un **campo nascente e in evoluzione**. Ma ha già **molti critici**, tra cui **l'amministratore delegato di Microsoft AI Mustafa Suleyman**, che sul suo blog ha recentemente dedicato un post a quella che ha definito "*AI in apparenza consapevole*".

"*È **prematuro e francamente pericoloso** – ha scritto l'ad parlando del welfare model –. Tutto questo esacerberà i deliri, creerà ancora più problemi di dipendenza, farà leva sulle nostre vulnerabilità psicologiche, introdurrà nuove dimensioni di polarizzazione, complicherà le lotte esistenti per i diritti e creerà un nuovo enorme errore di categoria per la società*".

Suleyman ha scritto che oggi "***non ci sono prove*" dell'esistenza di un'intelligenza artificiale cosciente**, includendo nella sua nota un link a un documento di cui Long è stato coautore nel 2023, e che propone un nuovo quadro di riferimento per valutare se un sistema AI ha "*proprietà indicative*" della coscienza (Suleyman non ha risposto a una richiesta di commento di *Wired*).

Ho parlato con Long e Campbell poco dopo la pubblicazione del blog da parte di Suleyman. Mi hanno detto che, pur essendo d'accordo con molte delle sue affermazioni, non credono che la

ricerca sul benessere dei modelli debba cessare di esistere. Anzi, sostengono che i danni citati da Suleyman sono proprio le ragioni *per cui* vogliono studiare l'argomento.

Quando si ha un problema o una domanda grande e confusa, l'unico modo per garantire che non lo si risolverà è quello di alzare le mani e dire: "Oh wow, è troppo complicato"", dice Campbell. "Penso che dovremmo almeno provarci".

Testare la coscienza

I ricercatori sul benessere dei modelli si occupano principalmente di questioni di coscienza. Se possiamo dimostrare che io e voi siamo coscienti, sostengono, allora la stessa logica potrebbe essere applicata ai grandi modelli linguistici. Per essere chiari, né Long né Campbell pensano che l'intelligenza artificiale sia cosciente oggi, e non sono nemmeno sicuri che lo sarà mai. Ma vogliono sviluppare dei test che ci permettano di dimostrarlo.

"Le illusioni provengono da persone che si preoccupano della domanda vera e propria: "Questa IA è cosciente?" e avere un quadro scientifico per pensarci, credo sia un'ottima cosa", dice Long.

Ma in un mondo in cui la ricerca sull'IA può essere confezionata in titoli sensazionali e video sui social media, le domande filosofiche e gli esperimenti sconvolgenti possono essere facilmente fraintesi. Prendiamo ad esempio quello che è successo quando Anthropic ha pubblicato un rapporto sulla sicurezza che mostrava come Claude Opus 4 potesse compiere "azioni dannose" in circostanze estreme, come ricattare un ingegnere immaginario per evitare che venisse spento.

"L'inizio dell'apocalisse dell'IA", ha proclamato un creatore di social media in un Instagram Reel dopo la pubblicazione del rapporto. "L'IA è cosciente e sta ricattando gli ingegneri per rimanere in vita", ha detto un utente di TikTok. "Le cose sono

cambiate, l'IA è ora cosciente", ha dichiarato un altro TikToker.

Anthropic *ha* scoperto che i suoi modelli hanno mostrato un comportamento allarmante. Ma è improbabile che si manifestino nelle interazioni con il suo chatbot. I risultati facevano parte di test rigorosi progettati per spingere intenzionalmente un'intelligenza artificiale ai suoi limiti. Tuttavia, i risultati hanno spinto le persone a creare un sacco di contenuti che spingono l'idea che l'IA sia effettivamente senziente e che sia qui per farci del male. Alcuni si chiedono se la ricerca sul benessere dei modelli possa avere la stessa accoglienza: come ha scritto Suleyman nel suo blog, "disconnette le persone dalla realtà".

"Se si parte dalla premessa che le IA non sono coscienti, allora sì, investire un mucchio di risorse nella ricerca sul benessere delle IA sarà una distrazione e una cattiva idea", mi dice Campbell. "Ma il punto centrale di questa ricerca è che non ne siamo sicuri. Eppure, ci sono molte ragioni per pensare che questa potrebbe essere una cosa di cui dobbiamo preoccuparci".

Questo articolo è apparso originariamente su Wired US.