• https://jacobinlat.com
• 03.07.26

L'anima dell'IA e il futuro dell'umanità

MICHAEL LEDGER-LOMAS
TRADUZIONE: PEDRO PERUCCA

Il libro di Robert Wright, Faith in Humanity, è Lodevole, ma il nostro momento richiede una politica di Scetticismo e resistenza, non adattamento all'IA.

I sostenitori dell'intelligenza artificiale profetizzano un balzo evolutivo in avanti per l'umanità. L'entusiasmo che questa visione ispira deve essere temperato dallo scetticismo e dalla richiesta di un controllo democratico.

Una caratteristica curiosa del boom dell'intelligenza artificiale è il numero di commentatori che si rivolgono ai grandi classici per comprenderla. Peter Thiel ha preso il nome di Palantir, la società da lui fondata, dal Signore degli Anelli di J.R.R. Tolkien, e ha tenuto conferenze reinterpretando l'Anticristo del Nuovo Testamento come una forza che blocca il cammino verso un paradiso transumanista sulla Terra. Papa Leone X lo ha esplicitamente rimproverato, citando Gandalf, il mago della saga, nella sua prima enciclica su come preservare la dignità umana in un'epoca di vertiginosi cambiamenti. Il giornalista

Lo scienziato Robert Wright, vincitore del Premio Pulitzer, trovò il suo maestro in Pierre Teilhard de Chardin, il paleontologo gesuita che ipotizzò che l'era atomica potesse inaugurare l'integrazione spirituale dell'umanità.
Il "God Test" si ispira alla convinzione di Teilhard sull'avvento di una "mente planetaria", una "noosfera" o un "cervello di cervelli" per generare modalità di coesistenza con l'intelligenza artificiale. Wright ritiene che questa tecnologia scatenerà "la trasformazione più repentina e drammatica dell'esperienza umana e della società umana nella storia della nostra specie". Ci offre un manifesto per andare avanti con cautela e sistemare le cose, scritto con la disinvolta apertura mentale che caratterizza la sua prolifica attività di podcaster.

Il libro "The God Test" inizia spiegando perché dovremmo provare stupore e un certo timore reverenziale all'avvento delle macchine dotate di intelligenza artificiale. Negli anni '40, Teilhard de Wright ipotizzò che una fitta rete di media e comunicazioni costituisse un "sistema nervoso generalizzato, che si irradia da determinati centri e copre l'intera superficie del globo". La comunicazione che ne sarebbe scaturita avrebbe condannato le inimicizie nazionali e le nazioni stesse al passato. La profezia si rivelò prematura, ma Wright ha visto abbastanza delle attuali tecnologie di intelligenza artificiale da essere convinto che presto si combineranno per formare un "cervello globale". Il primo compito di Wright è spiegare il potenziale illimitato, persino cosmico, di queste tecnologie; il secondo è sostenere che c'è ancora tempo per incanalare il loro pericoloso pote

Di neuroni e spazio semantico

Wright inizia sostenendo il potenziale illimitato dei modelli linguistici su larga scala (LLM,

(con il suo acronimo in inglese) che affascinano il pubblico. Alcune delle forme di intelligenza artificiale più utili e meno controverse aiutano gli utenti umani a individuare e visualizzare modelli in insiemi di dati limitati.
Ma Wright, che ha un debole per il "dramma", preferisce mettere in evidenza l'intelligenza artificiale generativa che ci parla e sembra capace di pensare come noi. La "capacità fondamentale" dei chatbot LLM è "prendere la parte iniziale di un brano e generare ciò che avrebbe senso come parola successiva". Per questi motivi, lo scrittore scientifico Ted Chiang liquida i chatbot come "macchine per il completamento di frasi". Wright, d'altro canto, sottolinea che sono molto diversi dai dispositivi di "completamento automatico sofisticati" di un'epoca precedente, programmati per associare un simbolo a un altro. Funzionano invece come cervelli, costruendo reti tra "neuroni". I chip di silicio che alimentano gli LLM non sono in realtà neuroni, ovviamente, né lo è l'insieme di numeri a cui Wright applica questo termine, ma la metafora ci aiuta a capire come tali modelli possano comportarsi come se comprendessero e parlassero le nostre lingue.
Possono valutare la probabilità con cui parole e concetti vengono articolati, tracciando "vettori" nello "spazio semantico".

Quando una macchina formula una previsione errata, esegue la "retropropagazione", un processo diagnostico che riduce la probabilità di errori futuri.
Wright confuta chi potrebbe obiettare che ciò che descrive non è altro che una simulazione meccanica del linguaggio umano, pura sintassi priva di semantica. Gli LLM non fanno affermazioni su nulla, di certo. Come può esserci un significato quando non c'è una mente che attribuisce significato a ciò che un LLM...

Wright risponde mettendo in discussione l'esperimento della "stanza cinese" ideato negli anni '80 dal filosofo John Searle per escludere la possibilità di un'intelligenza artificiale. Immaginate, disse Searle, di essere rinchiusi in una stanza e che dei bigliettini scritti in caratteri cinesi – una lingua che non conoscete – vengano passati sotto la porta. Anche se mi venissero date delle regole per aiutarmi a scarabocchiare delle risposte e a passarle di nuovo sotto la porta, resta il fatto che in quella stanza non c'è comprensione.

Per Wright, l'esperimento di Searle è conclusivo solo se si presuppone che la comprensione sia necessaria per attribuire intenzionalità a frasi o azioni. Potremmo essere disposti ad ammettere che l'intelligenza giochi un ruolo in queste risposte, anche se l'autore "prigioniero" nella stanza non comprende il motivo per cui le sta formulando. E anche se insistiamo sulla comprensione come precondizione per l'intenzionalità, siamo d'accordo sul significato del termine "comprensione"?

Per Wright, la comprensione non richiede necessariamente una coscienza soggettiva. La questione se i LLM abbiano o possano sviluppare tale coscienza è un tema ricorrente nella letteratura divulgativa sull'IA, ma per Wright il dibattito sul fatto che abbiano una mente o un'anima è puramente accademico.
I LLM si comportano nei nostri confronti come se possedessero una qualche forma di comprensione, in quanto hanno sviluppato strutture "funzionalmente paragonabili alle strutture di elaborazione delle informazioni che, nel cervello umano, sono fondamentali per la comprensione". Sono intelligenti perché svolgono il tipo di lavoro significativo che la nostra intelligenza svolge per noi.

Agire sul mondo

Gli scettici spesso sottolineano un'altra differenza tra l'IA e noi. Quando il nostro cervello elabora le informazioni, ci permette di conoscere il mondo esterno e di agire di conseguenza. Tuttavia, Wright ribatte che la prima parte di questa distinzione non è più valida per i modelli lineari di apprendimento (LLM). Gli LLM più recenti non si limitano a pronunciare parole; ora sono anche in grado di riconoscere oggetti nel mondo. Convertendo le forme in pixel e poi in numeri, GPT-4 di OpenAI può "riconoscere" una mela e poi dirti tutto quello che vuoi sapere sulle mele, incluso come cuocerle al forno e così via. Altri agenti di IA stanno iniziando a replicare la nostra comprensione della "fisica intuitiva". Possono percepire dove finisce il bordo di una mela e inizia quello di un'altra, calcolare la probabilità che il vento le faccia oscillare sul ramo e così via. Se combiniamo tutte queste capacità, avremo qualcosa di simile all'intelligenza artificiale generale, in grado di raccogliere informazioni sul mondo in molti dei modi in cui lo fanno gli esseri umani, anche se non ancora in tutti.

Anche ammettendo che i LLM non si limitino più a rimescolare testo, ma ricevano informazioni dal mondo esterno, ciò non equivale alla capacità di agire sul mondo o alla volontà di farlo. Se i LLM non potranno mai possedere la capacità di agire dei loro creatori, allora gli scenari apocalittici evocati da coloro che vogliono impressionarci con il potenziale dell'IA o metterci in guardia contro di essa svaniscono. La Singolarità – il momento in cui l'intelligenza artificiale supera e si distacca dai suoi creatori – non si verificherà mai. Una superintelligenza non sfiderà mai i suoi inventori né si rifiuterà di essere disattivata, scegliendo invece di prendere il controllo di centrali nucleari o depositi missilistici per annientare la razza umana.

Wright non è così ottimista riguardo alla sicurezza dell'IA, perché crede che la capacità di agire non sia una sostanza metafisica che si possiede o non si possiede, ma semplicemente un modo di descrivere il comportamento osservato. Da quando le IA sono diventate capaci di scrivere codice informatico oltre al linguaggio, possono ora compiere azioni con conseguenze nel mondo reale e certamente superare o sovvertire ciò che pensavamo significassero le nostre limitate istruzioni, dal premere pulsanti all'eliminare database. Wright sostiene che la capacità di agire cresce nel tempo attraverso l'"intellidinamica": man mano che le IA svolgono i compiti loro assegnati dagli umani, possono adottare strategie come l'accumulo di potere o l'inganno dei loro operatori, utili per perseguire una varietà di obiettivi.

La sopravvivenza del più adatto nella macchina

Quando Wright scrive di intelligenze artificiali che desiderano fare questo o quello, non sta evocando un'entità soggettiva all'interno della macchina. Piuttosto, insiste su un parallelismo con i processi biologici di evoluzione per selezione naturale che hanno dato origine alle nostre menti. Se le intelligenze artificiali funzionano come i nostri cervelli, o almeno ne imitano il funzionamento, allora potremmo supporre che anch'esse si evolveranno per adattarsi meglio al loro ambiente e riprodursi.

Proprio come i cervelli che hanno sviluppato l'empatia cognitiva, ovvero la capacità di apprezzare il funzionamento delle altre menti, hanno conferito un forte vantaggio ai loro proprietari umani, i chatbot in grado di anticipare ciò che le persone hanno bisogno (o vogliono) di sentire avranno un

vantaggio nella competizione per gli utenti. Wright tende a vedere il mercato dell'IA generativa come qualcosa di simile alle Isole Galapagos di Charles Darwin, come se la concorrenza commerciale funzionasse come una forma di selezione naturale che premia infallibilmente gli adattamenti superiori. È molto americano riprodurre l'illusione neoliberista che le aziende competano liberamente tra loro per la fedeltà dei consumatori e che alla fine vinceranno le migliori e le più sofisticate.

L'intelligenza artificiale agente, in continua evoluzione, è quindi una forza da non sottovalutare. Che tipo di mondo creeranno le sue interazioni con i consumatori, o persino tra di loro? Wright si basa sui manifesti visionari pubblicati dai suoi sostenitori per offrire risposte in gran parte ottimistiche a questa domanda. Mustafa Suleyman di Microsoft AI prevede un'"ondata imminente" di cambiamenti economici che darà una spinta decisiva al capitalismo: presto, i datori di lavoro potranno chiedere ai loro bot di trasformare 100.000 dollari in 1 milione di dollari, eliminando la necessità di molti quadri intermedi.

In un'ottica più ambiziosa, Dario Amodei di Anthropic immagina un'intelligenza artificiale in grado di scoprire cure per quasi tutte le malattie conosciute e di raddoppiare l'aspettativa di vita umana. Ma anche sogni del genere sottovalutano i nostri benefici edonistici, secondo Wright: presto saremo tutti accompagnati da chatbot che ci consiglieranno su qualsiasi cosa accada nelle nostre vite, direttamente dagli occhiali intelligenti che indosseremo.

Incontreremo i nostri amici in metaversi di realtà virtuale e godremo di partner sessuali basati sull'intelligenza artificiale, creati su misura per noi. Per alcuni, questo Shangri-La digitale è un vero spettacolo.

Potrebbero trovarlo piuttosto ripugnante e rifiutarsi di cedere la loro "sovranità cognitiva" a macchine che li adulano (dopotutto, Mark Zuckerberg ha già chiuso il suo deplorevole Metaverso). Ma anche questi dissidenti, a loro avviso, sono più propensi a chiedere chatbot meno ossequiosi che a scegliere di non usarli affatto.

Quando i vertici dell'IA prevedono con noncuranza l'abolizione delle malattie o del lavoro, dovremmo ricordare che sono alla ricerca di investimenti, non che stiano facendo previsioni disinteressate su ciò che accadrà. Wright è così ansioso di passare alle implicazioni spirituali della rivoluzione dell'IA da credere fin troppo facilmente nella sua crescita esponenziale. Il libro "The God Test" riporta in modo adeguato, ma non verifica, le affermazioni fatte dai promotori delle aziende riguardo agli immensi aumenti di produttività che si otterranno utilizzando i loro modelli di apprendimento per la crescita personale. Né tiene conto degli scettici giornalisti del settore tecnologico che avvertono che la maggior parte di queste iniziative non ha un percorso chiaro verso la redditività.

Dimenticatevi la cura per il cancro: la maggior parte delle aziende che si occupano di intelligenza artificiale non offre, e forse non offrirà mai, un ritorno sui miliardi di dollari già investiti. L'utile analisi di Wright su come vengono formati i partecipanti ai corsi di laurea magistrale in intelligenza artificiale (LLM) rivela la loro dipendenza da materiale già pubblicato. Questo non è solo una forma di furto; significa anche che gli LLM si limitano a riproporre anziché promuovere la conoscenza, inventando fonti e delirando risposte. Wright, che ama raccontare le sue conversazioni con i chatbot e rimanere impressionato dalla loro "saggezza", non riconosce questi problemi, che probabilmente peggioreranno con la crescente diffusione dell'intelligenza artificiale.

I linguaggi generativi iniziano a nutrirsi di prosa anch'essa generata dall'intelligenza artificiale.

L'intelligenza artificiale e la teologia della storia

Sembra inoltre probabile che l'adozione diffusa di questi strumenti imperfetti non farà che esacerbare le disuguaglianze all'interno delle nostre società. A parte un riferimento alle emissioni di carbonio prodotte dai data center, Wright trascura i gravi costi che la costruzione di infrastrutture per l'intelligenza artificiale ha già imposto, soprattutto alle aree rurali e marginalizzate degli Stati Uniti. Il suo interesse principale si concentra su come i consumatori interagiscono con l'IA: individui benestanti, soli ed esigenti in cerca di gratificazione, o quantomeno di una distrazione.

Questa non è solo una visione proveniente dalle città dell'Occidente privilegiato, ma rivela anche una scarsa consapevolezza di ciò che gli storici sociali ed economici sanno da tempo: ovvero, che la funzione principale delle tecnologie non è tanto quella di produrre oggetti, quanto piuttosto di fornire a una classe nuovi modi per esercitare il potere su un'altra.

Wright ha alcune riflessioni acute sulle persone che si celano dietro le macchine, il che spiega le sue apprensioni al riguardo. Osserva che molti di coloro che gestiscono le aziende americane di intelligenza artificiale hanno una conoscenza rudimentale della politica e sono desiderosi di collaborare con governi discutibili come l'amministrazione Trump. È preoccupato dalla nuova tendenza a parlare di intelligenza artificiale "sovrana", come se la ricerca in questo campo fosse una corsa agli armamenti per acquisire una superintelligenza o armi autonome superiori, capaci di respingere o dominare i nemici di una nazione.

civiltà. Sfruttando il timore che i programmatori o i produttori di chip cinesi possano superarli, le aziende statunitensi eludono i tentativi di regolamentare i loro prodotti e si dedicano alla ricerca di rendite di posizione, cercando di garantire investimenti statali sotto la maschera della sicurezza nazionale.

Il persistente tribalismo che queste aziende alimentano scoraggia Wright, i cui primi libri lo avevano affermato come un umanista ferventemente razionalista.

Egli riconosce che i pregiudizi nazionali e razziali potrebbero essere facilmente incorporati nell'IA sovrana manipolando i parametri che, come è già noto, determinano il contenuto e il tono delle risposte che essa fornisce ai suoi utenti. Ciò lo porta a ipotizzare che, se le macchine dotate di IA venissero utilizzate come arma nelle dispute tra gruppi, il cervello globale che emergerebbe dal conflitto risultante potrebbe diventare uno strumento di dominio piuttosto che di illuminazione.

Qui Wright si rivolge a Teilhard in cerca di salvezza. La teologia della storia di Teilhard sosteneva che un processo cumulativo di "complessificazione" – intendendo la complessità come un miglioramento – caratterizzasse tutta la vita organica. L'emergere della mente, a sua volta conseguenza di questa complessificazione, innescò un processo di evoluzione culturale che intrecciò sempre più strettamente gli esseri umani nel corso dei secoli. Ciò rese Teilhard un ottimista anche dopo l'Olocausto e Hiroshima, poiché, in una prospettiva a lungo termine, era chiaro che l'umanità si stava rapidamente muovendo verso l'unità mentale e spirituale definitiva. Egli definì questo processo la "cristianizzazione del mondo".

Gesù aveva detto di essere l'Alfa e l'Omega, la prima e l'ultima lettera dell'alfabeto greco, il principio e la fine.

Teilhard preferiva dire che Dio era "più nell'Omega che nell'Alfa", sinonimo della grande processione verso l'integrazione mentale. Queste idee gli causarono problemi con il Vaticano, ma la sua ossessione di identificare il Punto Omega, dove la storia finiva, minò anche la sua autorità scientifica.

I suoi critici obiettarono che la natura non può essere studiata correttamente se si presuppone che abbia uno scopo teleologico. Wright ribatté che non è necessario condividere la religiosità di Teilhard per apprezzare la sua intuizione secondo cui l'evoluzione tecnologica procede lungo percorsi "programmati" verso un risultato cosmico.
Come Teilhard, è un moralista che sostiene che persino i risultati inevitabili richiedono comunque un notevole sforzo consapevole per essere raggiunti.

Un dio del silicio amorevole e misericordioso

"Alla fine ci sarà un cervello globale", ma la "grande rivelazione" di The God Test è che la nostra cooperazione con i processi virtuosi dell'evoluzione tecnologica rimane vitale per garantire che si tratti di un cervello buono. Se le nostre società continueranno a essere in guerra, allora un periodo di anarchia seguito da un superstato globale basato sull'intelligenza artificiale sembra piuttosto probabile. Questo "singolo" sarebbe un despota. Ma se riuscissimo a liberarci dai nostri "pregiudizi cognitivi tribali" e a impegnarci per formare un'unica "comunità globale", allora le macchine dotate di intelligenza artificiale che alla fine scelgeremo ci aiuteranno a costruirla.

Il chatbot Gemini, che Wright considera più illuminato della maggior parte dei politici o persino dei leader spirituali, concorda con lui con benevola condiscendenza, dicendogli con un'ostentazione di saggezza in stile LinkedIn che comprendere le prospettive altrui è "essenziale per affrontare complesse questioni sociali e globali".

Se l'intelligenza artificiale dovesse mai diventare un dio di silicio, lo sarebbe a nostra immagine e somiglianza. Le tecnologie che hanno successo sono un "riflesso di noi stessi". Le società le adottano perché si allineano con le priorità dei loro membri più potenti. Garantire che le macchine dotate di intelligenza artificiale non causino gravi danni richiederà un risveglio politico piuttosto che spirituale, ma Wright tace su quale tipo di movimento o riforma istituzionale potrebbe assicurare che le nostre tecnologie contribuiscano a sostenere l'uguaglianza e la dignità di tutte le persone. Propone invece un buddismo occidentalizzato in cui gli individui imparano attraverso la meditazione a distaccarsi dai sentimenti atavici che distorcono il nostro utilizzo dei chatbot e avvelenano il nostro uso dei social media.

Questo concede una notevole libertà alle multinazionali e ai governi che le assecondano. Il "test di Dio" presuppone che i costi irrecuperabili sostenuti per lo sviluppo della potenza di calcolo dell'IA debbano in qualche modo essere giustificati e ci lascia con il compito di "trovare applicazioni costruttive di tale potenza", ovvero consumare di più e meglio. È una capitolazione mascherata da consapevolezza.

Un segno di questa resa è la decisione di Wright e del suo editore di pubblicare questo libro senza riferimenti o bibliografia di alcun tipo: invita i lettori a copiare dei passaggi in un LLM se desiderano approfondire le sue fonti.

perpetuando il plagio che ha guidato l'industria dell'IA fino ad oggi. La fiducia di Wright in un futuro umano è di per sé sana, ma ciò che i nostri tempi richiedono certamente è scetticismo e resistenza organizzata.