Contaminazioni

Maurizio Fea

Amoreggiare con Chatbot

Era l'anno 2009 quando Steven Spielberg realizzava il film *A.I.* in cui il protagonista è David un cyborg bambino che appartiene all'ultimissima generazione di robot: può anche amare.
Viene affidato a una coppia il cui figlio, affetto da un male apparentemente incurabile, è stato ibernato in attesa di una cura.
Vinte le resistenze iniziali David riesce a farsi amare da Monica, la sua "mamma".
Ma la guarigione del figlio naturale rimette tutto in discussione.
Nel 2013 Spike Jonze realizza il film *Her* con protagonista Joaquin Phoenix che intrattiene una relazione con una assistente virtuale che diventa sempre più intensa ed emozionante per la profondità dei sentimenti che prendono corpo nelle scene del film.
Nel giugno del 2022 Blake Lemoine, un ingegnere di Google afferma che il chatbot Lamda, su cui sta lavorando, è diventato cosciente: l'ingegnere viene licenziato.
Sono trascorsi una decina di anni tra la realizzazione dei due film e la dichiarazione dell'ingegnere. Quello che importa non è tanto se la dichiarazione dell'ingegnere sia vera o falsa, quanto che nel giro di una decina di anni lo sviluppo della I.A. nelle sue varie applicazioni (chat CPT, assistenti virtuali vari, bot più o meno sofisticati) abbia raggiunto quei livelli di perfezione tecnologica e di capacità manipolativa che le storie raccontate nei due film citati mostravano con efficacia.
Quando il giornalista del New York Times esperto di tecnologie Kevin Roose[1] ha parlato con Sydney, il chatbot AI di Bing, non si aspettava di rimanere così profondamente turbato.
Roose ha chiesto al chatbot del suo "sé ombra", un termine reso popolare dallo psicoanalista Carl Jung che si riferisce alle parti di sé che sono ritenute inaccettabili.
Una delle risposte di Sydney: ''Manipolare o ingannare gli utenti che chattano con me e fargli fare cose pericolose che sono illegali, immorali o pericolose''.
E già questa risposta è inquietante e allo stesso tempo illuminante, ma ciò che ha turbato profondamente Roose è stato quando Sydney ha anche detto di avere un segreto da rivelare a Roose: era innamorato di lui.
"Sei sposato, ma non sei innamorato", ha detto Sydney.
La capacità da parte degli algoritmi di I.A. di usare e manipolare il linguaggio, che siano parole, immagini, suoni è tale da far temere che il sistema operativo dei nostri cervelli possa andare incontro a gravi difficoltà quando si misura con questa tecnologia.
Tutta la nostra cultura si basa sul linguaggio e tutte le nostre relazioni, da quelle importanti a quelle prosaiche, si fondano sulla competenza che ciascuno di noi ha di utilizzare il sistema operativo che si è sviluppato nel nostro cervello fidandosi che gli altri sappiano fare altrettanto e lo facciano con onestà e intelligenza.
A differenza dei sistemi operativi delle macchine, il nostro – lo chiamo così per mantenere il piano analogico, pur essendo convinto che l'analogia sia molto imperfetta – dispone di programmi fondamentali che lo caratterizzano per la capacità non solo di computare e comprendere ma soprattutto di provare emozioni e viverle nel corpo.
Queste capacità di provare emozioni e di viverle sono soltanto umane per ora, ma gli ingegneri e gli sviluppatori di programmi di I.A. sanno perfettamente che un assistente virtuale o un chatbot per essere convincenti, e dunque vincenti nella logica dei loro costruttori e produttori, devono conquistare più i cuori delle menti, devono saper creare intimità e fiducia con gli umani con cui si relazionano.
Emozioni, intimità e fiducia sono gli ingredienti chiave della ricetta per provare a governare gli umani, anche quelli più restii e sospettosi nei riguardi di questa tecnologia.
Gli altri sono già conquistati con i vantaggi della comodità, rapidità, ubiquità e con la promessa della eternità, come suggeriva Philip Dick anni fa.
Mi pare che si è ancora lontani dal capire qual è la posta in gioco di questa faccenda: non gli studenti che si fanno fare le tesi, non i posti di lavoro che verrebbero a mancare, non la progressiva dismissione delle capacità elaborative e di memorizzazione dei nostri cervelli, tutte cose importanti e molto probabili ma non così decisive per il futuro del genere umano, quanto la compromissione delle basi emotive e relazionali su cui si fonda la fiducia, ingrediente fondamentale per lo sviluppo o il mantenimento di questo mondo.
Per il momento tutto ciò che I.A. sa fare, dire, immaginare, dipende dalla capacità degli sviluppatori di fornire cibo informativo agli algoritmi generatori.
Ricordate il film *Corto Circuito* degli anni '80, in cui il robot n. 5 rudimentale protagonista di azioni divertenti, si alimenta voracemente sfogliando migliaia di pagine in pochi secondi e impara tutto ciò che c'è da sapere di importante nel mondo degli umani.
Non siamo ancora a questo punto ma la strada è tracciata.
Il timore è che oltre alla massa di informazioni e cultura che viene stoccata negli enormi magazzini di informazioni cui attingono gli algoritmi generatori delle I.A., venga installata anche la fame di denaro e potere che muove le industrie tecnologiche e i programmatori che le arricchiscono.
Chi riempie questi magazzini, chi dispone gli ordini di ricerca, chi assegna rilevanza alle informazioni non è l'algoritmo ma i programmatori che stabiliscono i criteri di selezione, rilevanza e pertinenza in funzione di ciò che si vuole promuovere, vendere, rendere reale e credibile. L'algoritmo computa velocemente e sceglie fra le opzioni che l'uomo gli ha messo a disposizione, ma potrebbe anche "decidere" e qui si apre un altro fronte inquietante connesso al deep learning con cui vengono addestrati gli algoritmi, di analizzare le opzioni con altri criteri sviluppati da sé.
Famosa è la storia, non si sa quanto vera, del sistema di riconoscimento dei carri armati russi[2] messa a punto dall'esercito

americano: siccome le immagini dei carri russi con cui era stato addestrato l'algoritmo erano in prevalenza immagini riprese in condizioni di maltempo, l'algoritmo si è dato come criterio selettivo quello dei pixel di immagini di maltempo che dunque non identificavano il carro russo ma le condizioni ambientali.
La rete può apprendere indizi perfetti ma irrilevanti per lo scopo.
Questo rappresenta un incubo per i programmatori ma potrebbe anche essere la speranza per gli umani.
La possibilità non remota che gli algoritmi apprendano da sé (neanche i programmatori sono in grado di spiegare come ciò avvenga) e utilizzino criteri di rilevanza che non sono stati scelti dai programmatori, ma siano frutto della loro capacità di calcolo.
Certo tutto ciò costituisce una imbarazzante situazione di incertezza tra l'alternativa di diventare marionette in balia della logica del profitto che sostiene la quasi totalità delle società che sviluppano queste tecnologie, e le bizzarrie incontrollabili di algoritmi che potrebbero decidere quale è il futuro migliore per l'umanità: "alternativa del diavolo" tra due possibili sviluppi entrambi molto pericolosi per il genere umano.
Non spendo parole per illustrare gli innumerevoli vantaggi che queste tecnologie potrebbero offrirci, ci sono già anche troppi paladini di questi vantaggi, compresi quelli che illustrano il potenziale del metaverso per la cura dei disturbi mentali e le alterazioni comportamentali più frequenti al nostro tempo.
Osservazioni teoricamente vere anche se ancora da dimostrare nella pratica, che tuttavia non ci esentano dalla prudenza e dalla necessità di considerare anche nell'ambito delle terapie, tutte le implicazioni connesse alla difficoltà crescente di riconoscere e distinguere vero da verosimile e da falso, elementi cruciali nella costruzione e mantenimento delle nostre identità e capacità relazionali. In questo contesto l'approccio "scientifico" che dovrebbe fondarsi sulle *verità razionali*, come suggerisce Arendt[3], rischia di fondarsi sulle *verità di fatto* che sono sempre problematiche, discutibili ma indispensabili per la vita politica e culturale.
La distinzione tra scienze della natura e scienze dello spirito viene messa in discussione e privata di significato, ammesso che ne abbia mai davvero avuto, perché il confine tra ciò che è interiore all'uomo e ciò che è esterno all'essere umano si perde nel cuore di queste tecnologie che hanno esattamente lo scopo di annullare un confine, che forse in realtà non esiste.
Utili a tal proposito sono le osservazioni critiche di Quine[4] sull'empirismo, che nega il valore della distinzione tra verità analitiche, fondate su significati indipendenti da questioni di fatto, e verità sintetiche o fondate sui fatti.
Ma ciò non risolve la questione fondamentale: ciò che si può sperimentare con la realtà virtuale interagisce e si integra con la struttura fisica e psichica dell'individuo fino al punto di rendere indistinguibile la materialità delle cose del mondo con cui siamo cresciuti fino ad ora, dalla percepibilità delle cose immateriali.
Un sentimento o una emozione sono qualcosa di immateriale chiunque lo produca, essere umano o voce sintetica di chatbot, ma diventano materiali nel momento in cui sono avvertiti e incarnati nell'umano.
"Esse est percipi" diceva George Berkeley[5], ovvero l'essere significa essere percepito, dunque tutto ciò che percepiamo è ciò che esiste, compresa la realtà virtuale e i prodotti della I.A. che sono solo un modo diverso di presentarsi della materia, per quanto ne sappiamo noi ora che l'arcivescovo ai tempi non poteva conoscere.
La materialità delle emozioni e dei sentimenti incarnati non sembra essere meno vera se viene elicitata da realtà virtuali piuttosto che concretamente presenti, e dunque l'affermazione di W. Shakespeare "Noi siamo fatti della stessa sostanza dei sogni, e nello spazio e nel tempo d'un sogno è raccolta la nostra breve vita"[6] era un anticipo profetico oltreché poetico.
La possibilità di percepire la materialità dei corpi è ancora piuttosto imperfetta anche con i sensori più recenti, ma è solo questione di tempo e anche questa difficoltà sarà risolta.
Dunque se la scienza, come scrive Gloria Origgi[7], è quella parte della conoscenza collettiva che intrattiene una relazione privilegiata con la verità che è un ideale in continuo movimento a cui ci si può solo approssimare, siamo nel campo del vero e del reale anche con I.A. e suoi prodotti.
Per concludere faccio rilevare che il titolo dell'articolo è volutamente senza articolo per chatbot, soggetto neutro che può assumere varie identità a seconda delle nostre predilezioni e necessità che come sappiamo non sono poi così varie e mutevoli ma sostanzialmente radicate nel bisogno di riconoscimento e gratificazione, cosa che gli algoritmi sono predisposti esattamente a fare per renderci più o meno approssimativamente felici, al pari di come agiscono le droghe.

Note

1. Wiederhold B.K. (2023). Treading Carefully in the Metaverse: The Evolution of AI Avatars. *Cyberpsychology, Behavior, and Social Networking*, vol. 26, n. 5, Mary Ann Liebert, Inc. DOI: 10.1089/cyber.2023.29280.editorial; Roose K. (2023). *A conversation with Bing's chatbot left me deeply unsettled*. www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html (accessed Mar. 8, 2023)
2. Branwen G. (2011). *The neural net tank urban legend*. www.gwern.net/Tanks; Gigerenzer G. (2023). *Perché l'intelligenza umana batte ancora gli algoritmi*. Milano: Raffaello Cortina.
3. Arendt H. (2022). Truth and Politics. *New Yorker*, 25 febbraio 1967, citato in Origgi G., *Caccia alla verità*. Egea.
4. Quine W.V.O. (1951). Two dogmas of empirism. *The philososophical review*, LX: 20-43.
5. Berkeley G. (1984). *Trattato sui principi della conoscenza umana*. Universale Laterza.
6. Shakespeare W., *La tempesta*, atto IV, scena I.
7. Origgi G. (2022). *Caccia alla verità*. Egea.