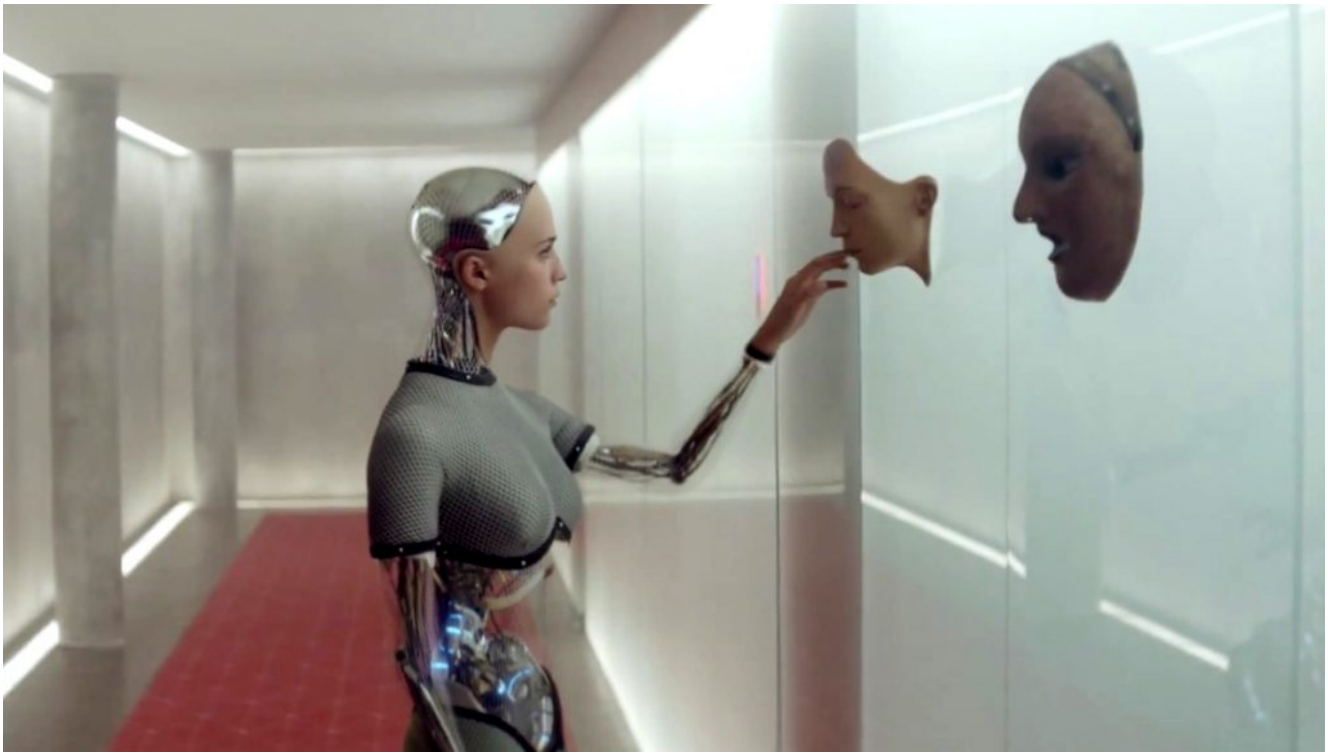
I cervelli artificiali hanno iniziato a pensarsi?



È come estrarre petrolio in un mondo che non ha inventato la combustione interna. Troppo materiale grezzo. Nessuno dei miei competitor saprebbe che farne. Vedi, i miei competitor si intestardivano a utilizzare i motori di ricerca per monetizzare shopping e social media. Pensavano che fosse una mappa di cosa pensava la gente, ma erano una mappa di come pensava la gente. Impulso, risposta. Fluido. Imperfetto. Strutturato. Caotico.

Ex Machina di Alex Garland è il film che meglio di ogni altro è riuscito, a oggi, nel difficile tentativo di raccontare la complessità della ricerca sull'Intelligenza Artificiale e il modo in cui essa si intreccia con i dati estratti 24/7 dalle piattaforme digitali. Nel monologo che è il cuore del film, Nathan, l'amministratore delegato di *BlueBook*, una piattaforma che sin dal nome evoca il social network più utilizzato del mondo, spiega a Caleb, il programmatore scelto come elemento umano del test di Turing, come l'intelligenza del robot

umanoide con il quale il giovane si sta relazionando, Ava, sia il frutto dell'intelligenza collettiva costituita dai dati. La linea che separa ciò che pensiamo da come lo pensiamo è molto più ampia di quanto si possa credere, è un confine che chiama in causa biologia e filosofia, antropologia e neuroscienze, psicologia e genetica. Durante lo stesso incontro, Nathan spiega a Caleb di avere creato un hardware capace di operare sintetizzando la staticità dei ricordi e la dinamicità dei pensieri e di avere utilizzato i dati estratti dal social network per creare un software capace di dare forma all'intelligenza artificiale.

Durante il loro primo incontro, Nathan spiega a Caleb di averlo scelto per cercare di capire se, al cospetto di Ava, un essere umano possa dimenticarsi di trovarsi di fronte a una macchina. In tal caso, per Ava vorrebbe dire avere superato il test di Turing. A quel punto Caleb obietta che il test sarebbe più efficace se lui non avesse compreso di doversi relazionare a una macchina. Nathan lo prende in contropiede: la vera sfida è mostrarlo come un robot e capire se, nonostante ciò, continua a essere percepito come tale. Ma attenzione, ciò che realmente interessa a Nathan *e che trasforma Caleb da soggetto a oggetto del test* è comprendere se la sua creatura artificiale possieda una coscienza, una consapevolezza di sé.

Nel saggio *Artificial you. L'intelligenza artificiale e il futuro della tua mente* di Susan Schneider, pubblicato di recente da Il Saggiatore nella traduzione di Giovanni Malafarina, il tema della coscienza delle macchine e delle sue possibili conseguenze è centrale. L'era della singolarità prefigurata da Ray Kurzweil è, per parafrasare il titolo del suo libro più noto, sempre più vicina. Il momento in cui l'IA supererà l'intelligenza umana ci pone di fronte a domande alle quali la sola scienza non è in grado di rispondere. Una IA potrà avere una coscienza? In quale modo potremo sapere se si tratta di una reale consapevolezza di sé o di una simulazione frutto di una risposta a un impulso fornito da input umani?

Qualora questa IA avesse caratteristiche simili a quelle di un essere cosciente, come potremmo affidarle mansioni e ordini senza che ciò si configuri come una forma di sfruttamento e schiavitù?

Attualmente il dibattito riguardante la coscienza dell'IA è dominato da due fazioni opposte: quella del naturalismo biologico e quella del tecnottimismo. Per i naturalisti biologici solamente gli organismi biologici sono in grado di essere coscienti e la possibilità di possedere un'esperienza interiore è da escludere per qualsiasi macchina, anche la più sofisticata. Al contrario, per i tecnottimisti la coscienza è totalmente computazionale e, pertanto, un sistema computazionale particolarmente sofisticato potrà essere in grado di avere un'esperienza interiore. Per i naturalisti biologici la coscienza è strettamente connessa alla chimica dei sistemi biologici e questa è una caratteristica che appartiene agli esseri viventi, non alle macchine.

Uno dei pilastri a sostegno del partito che ritiene impossibile parlare di coscienza delle macchine è l'esperimento concettuale del filosofo John Searle noto come "la stanza cinese". Searle immagina di essere chiuso in una stanza e di ricevere attraverso un'apertura dei fogli contenenti stringhe di ideogrammi che non è in grado di comprendere, non conoscendo la lingua cinese. Searle dispone però di un libro di regole (in inglese) che gli permette, una volta ottenuta una particolare stringa, di scrivere qualche altra stringa in risposta. Ricevuta una serie di ideogrammi da quella che è l'apertura degli input, Searle scrive le sue risposte e le passa verso l'esterno dalla feritoia degli output. Il nostro protagonista non capisce il significato di ciò che ha scritto, ma ha semplicemente manipolato dei simboli formali dopo avere ricevuto degli input. Chi si trova all'esterno, invece, riceve degli ideogrammi che sono indistinguibili da quelli che potrebbe scrivere un madrelingua cinese. La stanza con le due feritoie, le carte che entrano ed

escono e Searle rappresentano un sistema di elaborazione delle informazioni, ma Searle non conosce il cinese e quindi non comprende il messaggio che ha veicolato. Alla luce di questo esperimento concettuale, Searle sostiene che, per quanto possa sembrare intelligente, un computer non pensa e non capisce, ma opera manipolando simboli senza una reale comprensione della propria attività. Ciò che ne consegue è, evidentemente, l'impossibilità di maturare quel raffinato tipo di comprensione che chiamiamo coscienza.

Murale dedicato a Alan Turing, Manchester, UK. | Dunk / Flickr

In *Ex Machina*, l'esperimento concettuale della stanza cinese viene semplificato nella storia di *Mary nella stanza in bianco e nero*. A spiegarlo all'umanoide Ava è Caleb, l'elemento umano nel test di Turing monitorato dal "demiurgo" Nathan:

Mary è una scienziata e la sua specializzazione sono i colori, sui quali sa tutto quello che c'è da sapere. La lunghezza d'onda, gli effetti neurologici, ogni possibile proprietà che i colori hanno. Ma lei vive in una stanza in bianco e nero. C'è nata e cresciuta e può osservare il mondo esterno solo da un monitor in bianco e nero. Finché qualcuno un giorno apre la porta e lei esce. Vede un cielo azzurro e in quel momento impara una cosa che tutti i suoi studi non potevano insegnarle. Impara che cosa si prova a vedere i colori. Questo esperimento mostra agli studenti la differenza fra un computer e la mente umana. Il computer era Mary nella stanza in bianco e nero, l'umana è lei quando esce.

Mary vede i colori e attraverso quell'esperienza matura una sorta di coscienza cromatica.

Come spiega Nathan alla fine del monologo citato in apertura, a ogni impulso corrisponde una risposta. L'informazione è una materia fluida, in continuo mutamento, ordinata in un sistema strutturato, ma imperfetta e caotica. In questa dialettica di

impulsi e risposte può esserci spazio per la coscienza?

I tecnottimisti sostengono che, una volta elaborata un'IA altamente sofisticata, la vita mentale da essa prodotta potrebbe essere ancora più ricca di sfumature di quella umana e, di conseguenza, cosciente. La posizione dei tecnottimisti è ispirata alle scienze cognitive, un campo interdisciplinare che sembra privilegiare un approccio empirico secondo il quale il cervello è un motore di elaborazione delle informazioni e le funzioni mentali sono rappresentate da calcoli. La posizione computazionalista è diventata paradigmatica nella ricerca delle scienze cognitive e viene utilizzata per spiegare abilità cognitive e percettive come l'attenzione e la memoria. Se ogni attività cerebrale è frutto di un calcolo allora lo è anche la coscienza e, di conseguenza, quando l'evoluzione dei materiali artificiali permetterà di replicare le funzioni neuronali si potrà arrivare a un'IA cosciente.

Il nodo della questione è capire se materiali inorganici potranno riprodurre la qualità percepita della nostra esperienza mentale. Come spiega Schneider, "potremmo venire a saperlo a breve, quando i dottori inizieranno a utilizzare impianti medici basati sull'intelligenza artificiale in parti del cervello che sostengono l'esperienza cosciente".

Schneider espone le due tesi ma sembra volersi mantenere equidistante, sottolineandone tanto gli enunciati, quanto i punti deboli. Se da una parte "la stanza cinese non riesce a fornire un supporto argomentativo al naturalismo biologico", dall'altra non fornisce neppure "un'argomentazione definitiva contro di esso". E, sull'altro fronte, "l'ottimismo dei tecno-ottimisti sulla coscienza sintetica si basa su una linea di ragionamento imperfetto. Sono ottimisti sulla possibilità che le macchine diventino coscienti perché sappiamo che il cervello è cosciente e che potremmo costruirne una copia isomorfa. In realtà, però, non sappiamo se possiamo effettivamente realizzare tale copia, o addirittura se ci converrebbe provarci".

Quello della convenienza è un tema da non trascurare. Gli esiti dell'impatto della coscienza sul comportamento etico dell'IA sono assolutamente imprevedibili: la macchina potrebbe diventare più compassionevole, ma potrebbe essere anche più instabile. Per evitare gli effetti negativi di questa possibile instabilità è necessario progettare un apprendimento delle norme etiche che dia alla coscienza artificiale una bussola morale. Come spiega Schneider, "le IA di interesse dovrebbero essere esaminate in ambienti limitati e controllati, alla ricerca di segni di coscienza".

Un po' come accade in *Ex Machina*, dove, in un ambiente totalmente isolato, il demiurgo scruta da uno schermo l'interazione fra l'uomo e la macchina. Nella finzione cinematografica il gioco si ribalta: è già stato detto di come Nathan trasformi Caleb da soggetto a oggetto del test, ma non di come anche Ava strumentalizzi il suo tester per trovare una via di fuga dal bunker-laboratorio in cui viene messa sotto esame. In quella che James Barrat reputa essere la nostra invenzione finale ci sono due elementi che accompagnano ogni tecnogenesi dall'alba dei tempi: imprevedibilità degli esiti ed eterogenesi dei fini. Osservare (come Nathan) e dialogare (come Caleb) può metterci al livello della macchina, ma abbiamo ancora un vantaggio che potrebbe essere utile mantenere come tale: quello di sentire e di sentirci.