Tristi macchine allo specchio

Uno degli aspetti più affascinanti dell'IA è il suo impatto sulla rappresentazione che abbiamo di noi stessi. Quanto più grande si fa la capacità di macchine artificiali di replicare funzioni cognitive umane di alto livello, tanto più irresistibile si fa l'immagine 'meccanistica' della persona umana, come risultante dall'interazione stratificata di una molteplicità di meccanismi sub-personali. Una macchina biologica. In questo breve articolo ripercorrerò alcuni aspetti di questi rispecchiamenti di macchina e natura, e dell'angoscia che provoca in molti (ma non in tutti).

1. Macchine e specchi

Uno degli aspetti più affascinanti dell'IA è il suo impatto sulla rappresentazione che abbiamo di noi stessi, della nostra mente e del nostro posto nel mondo. La comprensione delle nostre abilità cognitive 'naturali' è indissolubilmente intrecciata, in un rapporto di reciproco rinforzo, con la capacità di costruire meccanismi artificiali in grado di simularle, replicarle, e a volte potenziarle. Quanto più grande si fa la capacità di macchine artificiali di emulare funzioni cognitive umane di alto livello, e quanto più profonda si fa la comprensione dei loro meccanismi naturali, tanto più irresistibile diventa l'immagine 'meccanistica' della persona umana, come risultante dall'interazione stratificata di una molteplicità di meccanismi sub-personali. Una 'macchina' biologica.

È un irresistibile gioco di specchi. Cerchiamo di costruire macchine che sappiano fare le cose che sappiamo fare noi. Per riuscirci, dobbiamo anzitutto capire cos'è, di come siamo fatti, che ci permette di fare le cose che facciamo. Per poi replicarlo, facendo macchine modellate a nostra (più o meno approssimativa) immagine e somiglianza. Quanto più tentiamo e quanto più riusciamo – mai abbastanza, o forse già troppo –, tanto più il riflesso torna indietro sulla sua matrice, e la cambia: noi ci riconosciamo nella macchina, e ci vediamo come macchine. E questo è per molti (ma non per tutti) perturbante, angoscioso.

In questo breve testo, ripercorrerò – in modo inevitabilmente sommario, selettivo e drasticamente semplificato, ma, spero, non troppo distorto (un po' sì, e me ne scuso) – alcuni aspetti di questo processo di rispecchiamento tra macchina e natura che attraversa sia scienza e filosofia che senso comune, per arrivare a sfiorare, alla fine, l'angoscia che provoca in molti (ma non in tutti).

2. Dalla natura alla macchina e ritorno

L'intreccio tra natura e macchina ha accompagnato inestricabile le scienze cognitive sin dalla loro fondazione avvenuta, tra gli anni cinquanta e settanta del secolo scorso, attraverso l'incontro e le collaborazioni, via via sempre più strutturate, fra studiosi impegnati in ricerche nel campo dell'IA con psicologi, linguisti, neuroscienziati. Sin dall'inizio, esse sono state concepite come scienze della cognizione *naturale e artificiale*. Anche la riflessione filosofica che sin dall'origine forma parte integrante del progetto delle scienze cognitive

nasce e si sviluppa nel segno dell'intreccio tra natura e macchina, e del continuo rimbalzo dall'una all'altra[1].

L'esempio più ovvio è il contrasto fra architetture simboliche e architetture connessioniste, che ha costituito, a partire dagli ultimi decenni del secolo scorso, il principale tema di discussione in quest'ambito di studi. È un contrasto che riguarda tanto la mente artificiale quanto quella naturale, e che riverbera dall'uno all'altro piano. È utile richiamarne rapidamente gli aspetti essenziali[2].

Il contrasto ha opposto, anzitutto, due diversi modelli e programmi di ricerca in IA: uno, l'IA classica o simbolica, basato su operazioni di manipolazione di simboli sulla base di regole esplicite, e l'altro, l'IA connessionista, basato su informazione distribuita nei pesi delle connessioni di reti neurali artificiali. Entrambi i modelli, è appena il caso di notarlo, si ispirano ad aspetti della cognizione naturale.

L'IA classica si ispira ai sistemi della logica formale: notazioni simboliche e insiemi di regole 'sintattiche' per la composizione e trasformazione dei simboli (regole, cioè, che operano sulla sola forma dei simboli, indipendentemente dalla loro interpretazione), attraverso i quali la ricerca logica, a partire dalla seconda metà dell'ottocento, è riuscita a catturare la struttura normativa profonda del pensiero e del linguaggio naturale. Questo sforzo di formalizzazione della cognizione naturale si è trasformato in un progetto di meccanizzazione quando si è compreso che ogni serie ben definita di operazioni formali può essere svolta da una macchina simbolica semplicissima, la Macchina universale di Turing, e che una Macchina universale di Turing può essere realizzata da un computer elettronico digitale.

L'IA connessionista si ispira invece alla struttura di base dei sistemi neurali e alla loro caratteristica forma di apprendimento, la variazione della forza delle connessioni fra neuroni, o plasticità neurale. Questa stilizzazione della architettura del cervello si è trasformata in un progetto d meccanizzazione quando si è compreso come costruire reti neurali artificiali e come simularne la plasticità attraverso un appropriato algoritmo.

Ma il contrasto tra IA classica e IA connessionista ha anche avuto fortissimi riverberi sul piano della mente naturale, dando vita a due diverse raffigurazioni della sua struttura e del metodo appropriato per studiarla.

Il primo modello è la cosiddetta teoria rappresentazionale della mente. L'idea centrale è che la cognizione naturale possa essere interamente spiegata in termini di manipolazione di simboli governata da regole sintattiche. Il pensiero, in particolare, avrebbe struttura linguistica: pensare significa concepire frasi in 'linguaggio del pensiero', stringhe di simboli discreti composti e trasformati sulla base di regole sintattiche. Per quello che qui più conta, in questo modello la mente è raffigurata come strettamente analoga ad una IA classica: una macchina simbolica. Proprio, in virtù di questa corrispondenza, si ritiene, una IA classica può emulare con successo le caratteristiche della cognizione naturale. Per farlo, deve ricostruire il 'programma' appropriato: un insieme di regole equivalenti a quelle

[1] Per chi volesse approfondire, rinvio a Bechtel et al. 1998, la più bella ed esaustiva ricostruzione a me nota della gestazione, nascita e crescita delle scienze cognitive; Bermúdez 2017, una ricognizione molto informativa e aggiornata; Miller 2003, una panoramica sintetica e illuminante, offerta da uno dei pionieri delle scienze cognitive.

[2] Il lettore troverà una snella ma brillante introduzione alla questione e alle sue implicazioni per la comprensione delle menti naturali in Churchland & Churchland 2000.

seguite dalla mente naturale nello svolgimento del compito che si intende emulare. Questa ricostruzione si colloca ad un livello di indagine (quello della psicologia) completamente indipendente da quello relativo alla implementazione fisica del programma nel cervello (il livello delle neuroscienze). La mente naturale *è realizzata dal* cervello, ma *non è* il cervello.

Per il secondo modello, l'attività cognitiva naturale non consiste nella manipolazione di simboli secondo regole sintattiche, ma piuttosto in pattern di attivazione neurale modulata (fra l'altro) dalla forza delle connessioni sinaptiche. Per quello che qui più conta, in questo modello la mente è strettamente analoga ad una IA connessionista. È quest'ultima, e non l'IA classica, che può emulare con successo le caratteristiche della cognizione naturale – non ricostruendo un 'programma', ma riproducendo in modo sempre più fedele la struttura e le dinamiche cerebrali.

3. In fuga dalla macchina

Nell'esempio del paragrafo precedente, il rapporto fra mente naturale e artificiale è un rapporto di assimilazione: sia nella teoria rappresentazionale che in quella connessionista la mente naturale è trattata come una macchina biologica strettamente analoga alle macchine artificiali. Possono esservi certamente rilevanti differenze *quantitative* rispetto a cosa possono fare le macchine artificiali e le macchine biologiche, ma non vi è una rilevante differenza *qualitativa*: esse svolgono lo stesso tipo di funzioni, attraverso lo stesso tipo di strutture e di processi.

Ma la riflessione filosofica sul rapporto fra mente naturale e mente artificiale è stata, anche e soprattutto, impegnata nell'elaborare argomenti per *refutare* la assimilazione della natura alla macchina.

Questo rifiuto può assumere tratti diversi. In alcuni casi, può trattarsi semplicemente del rifiuto di assimilare le menti naturali a *tipi contingenti* di macchine artificiali, senza con ciò escludere che si diano o possano darsi macchine artificiali capaci di superare i limiti attuali ed approssimarsi sufficientemente alla cognizione naturale. A volte, può anche trattarsi della rilevazione dei limiti attuali dell'IA proprio in vista del loro superamento. Per esempio, un tipico argomento diretto dai teorici connessionisti contro l'IA classica faceva valere la sua incapacità di riprodurre alcune abilità cognitive basilari delle menti naturali, come l'immediato riconoscimento di pattern, proprio per rimarcare la necessità di sistemi come quelli dell'IA connessionista che, invece, eccelleva nel riconoscimento di pattern .

In altri casi, la reazione all'assimilazione prende i tratti di una vera e propria *fuga* dalla macchina – una difesa a oltranza dell'irriducibile specificità della mente naturale rispetto alla macchina artificiale che la incalza per assimilarla a sé. La strategia di difesa è di almeno due tipi. Chiamerò la prima, apparentemente più compiacente, strategia delle 'inimitabili macchine biologiche', e la seconda, più drastica, strategia delle 'macchine mai!'.

Un esempio del primo tipo di strategia è il celebre argomento della stanza cinese di Searle (1980). L'argomento è diretto contro alcune versioni dell'IA classica, e più precisamente contro la pretesa che una macchina che processa simboli non interpretati sulla base di regole sintattiche 'comprenda' i simboli stessi se solo è in grado, se interrogata, di dare regolarmente il tipo di risposta che verrebbe data da un agente umano che

comprendesse i simboli attribuendo ad essi l'interpretazione adeguata (test di Turing). Searle fa l'esempio di un uomo chiuso in una stanza. Chiamiamolo John. John riceve da una fessura bigliettini con domande scritte in cinese, lingua che non conosce, e risponde in cinese consultando un libro di istruzioni che specifica che simboli usare per rispondere ai simboli ricevuti. Ebbene, Searle argomenta, John non comprende le domande e le risposte in nessun senso di 'comprende', proprio perché, non conoscendo il cinese, non è in grado di interpretare i simboli. La capacità di comprensione richiede la capacità di connettere i simboli al mondo assegnandogli un significato ('intenzionalità'). Ma questa facoltà, Searle ritiene, è posseduta solo dai cervelli di organismi viventi. Searle non esita a riferirsi ai cervelli degli organismi viventi come 'macchine biologiche'. Si tratta però di macchine biologiche inimitabili, impossibili da contraffare artificialmente.

Nonostante l'indubbia finezza dell'argomento, la conclusione riguardo alle 'inimitabili macchine biologiche' è stata criticata da molti come una petizione di principio (si veda, per un esempio brillante e sintetico, il già citato Churchland & Churchland 2000). Sono d'accordo nella diagnosi, e vorrei aggiungere un ulteriore spunto. Cosa c'è nella comprensione di stringhe di simboli che manca a John rispetto al cinese? C'è, anzitutto, la capacità di associare immagini percettive complesse ai simboli. Supponiamo per esempio che davanti alla stanza, visibile a John, vi sia una mela poggiata su un tavolo, e che la domanda posta a John sia 'La mela è sul tavolo?'. Nella comprensione di questa domanda vi è (non solo, ma anche) la capacità di formare un'immagine percettiva di una mela sul tavolo e di associare al simbolo 'La mela' un'immagine percettiva dell'oggetto mela, al simbolo 'tavolo' un'immagine percettiva dell'oggetto tavolo, e al simbolo 'è sul' la relazione spaziale del trovarsi sopra (sul contenuto intenzionale come immagine percettiva v. Barsalou 1999). Se un agente avesse la capacità di formare queste immagini e di associarle ai simboli appropriati, non ci verrebbe naturale dire che quell'agente ha un certo grado di comprensione dei simboli? Se, ad esempio, John associasse le appropriate immagini percettive alle corrispondenti espressioni cinesi, non verrebbe naturale dire che John 'capisce almeno un po'' il cinese? Ma non vi è nessuna ragione per escludere che macchine artificiali possano formare immagini percettive e associarle a simboli. Possono. Certo, si potrebbe obiettare, nella comprensione di John c'è molto più che queste capacità. Vi sono anche svariate abilità complesse, sia extra che intra-linguistiche. Ma non vi è ragione per escludere che macchine artificiali non possano avere parte almeno di queste abilità. Se così fosse, la macchina 'capirebbe' in misura ancora maggiore. La differenza tra macchine biologiche e macchine artificiali comincia, così, a diventare una differenza quantitativa, e non qualitativa. Siamo macchine biologiche non del tutto inimitabili, in fin de conti.

È a questo punto che si apre la seconda, notissima, strategia di difesa. Le macchine, si argomenta, non hanno quello che i filosofi chiamano 'coscienza fenomenica', o stati 'qualitativi': l'esperienza peculiare, indefinibile e irriducibile, di vedere rosso quando vedo rosso, provare paura quando provo paura, visualizzare il numero tre quando visualizzo il numero tre, ecc. È questa la strategia delle 'macchine mai!'. Una macchina può (forse) avere una cognizione per svariati aspetti simile a quella naturale. Non solo può manipolare simboli secondo regole sintattiche, ma può anche formare immagini percettive ed associare ad esse simboli (interpretarli, comprenderli almeno un po'), può avere abilità complesse di vario genere, può persino avere una forma di coscienza, la cosiddetta 'coscienza-accesso'

(Block 1995) – grossomodo, la capacità di mantenere una informazione attiva nella memoria di lavoro così che possa essere oggetto di processi cognitivi di alto livello. Ma una macchina non può avere la più peculiare e misteriosa forma di coscienza, la coscienza fenomenica. Se si vuole, continua l'argomento, il termine 'macchina' si può estendere, oltre che a macchine artificiali, anche ad entità biologiche. Ma, nella misura in cui sono 'soltanto' macchine – e qui macchine significa: meri assemblaggi di *materia* – queste entità sono prive di coscienza fenomenica. È questo l'ultimo bastione di difesa dell'assimilazione alla macchina: la supposta irriducibilità della coscienza alla mera sostanza fisica.

4. Catturati dalla macchina

Vorrei adesso accennare brevemente ad una recente, importante corrente di filosofia della scienza che si auto-qualifica, orgogliosamente, come *neo-meccanicismo* (Glennan 2017; Craver & Tabery 2019).

Il neo-meccanicismo nasce come razionalizzazione del modello di spiegazione adottato dalle scienze che studiano organismi e processi cognitivi *naturali*: la biologia e le neuroscienze cognitive (Bechtel & Abrahmsen 2005; Craver 2007). Ma ambisce a proporsi come modello generale di spiegazione scientifica applicabile, e già applicato, nei domini più vari.

La nozione centrale per i filosofi neo-meccanicisti è quella di 'meccanismo.' Un meccanismo può essere definito, in modo molto ampio, come una struttura fisica composta di parti, organizzate in modo tale che le loro attività producano regolarmente un certo 'fenomeno', e cioè un pattern o evento osservabile. Molti meccanismi sono stratificati: le loro parti sono, cioè, a loro volta meccanismi, composti da altri meccanismi, su innumerevoli livelli di organizzazione.

Il tipo di meccanismi qui rilevanti sono i cosiddetti meccanismi 'mentali' (Bechtel 2008). In prima approssimazione, i meccanismi mentali possono essere definiti come meccanismi che regolano dinamicamente l'interazione tra un organismo vivente e l'ambiente (ovvero interazioni tra parti dell'organismo, funzionalmente connesse all'interazione con l'ambiente) attraverso il coinvolgimento del sistema nervoso centrale. Questa nozione di meccanismo mentale si applica, ovviamente, solo a sistemi cognitivi naturali. Ma non vi è alcuna ragione che impedisca di estendere la nozione fino ad includere qualsiasi meccanismo che svolga funzioni analoghe in entità artificiali. In ogni caso, ciò che adesso ci interessa sono proprio i meccanismi mentali naturali.

La nozione di meccanismo mentale è al centro di una concezione molto articolata della spiegazione dei fenomeni mentali. In estrema sintesi, la si può riassumere così. Spiegare un fenomeno mentale target significa mostrare come esso sia prodotto dalle operazioni di certi meccanismi mentali. Questo tipo di spiegazione ricomprende tipicamente almeno due livelli, un livello 'funzionale-omuncolare' e un livello neurale. Al livello funzionale-omuncolare, tipico dell'indagine di psicologia cognitiva, le operazioni sono descritte come se fossero svolte da omuncoli intelligenti (Dennett 1978, Lycan 1981), senza specificare quali siano le strutture e i processi fisici che le realizzano – ad esempio, il fenomeno della memoria può essere distinto in operazioni di immagazzinamento, recupero,

ecc.. A livello neurale, tipico dell'indagine delle neuroscienze cognitive, l'omuncolo è dissolto mostrando come le operazioni in questione possano essere svolte da strutture neurali sub-personali. (Il processo è circolare: l'analisi condotta a livello neurale può condurre a ridisegnare le operazioni rilevanti, o anche a modificare la descrizione iniziale del fenomeno.)

La concezione su descritta si accompagna, tipicamente, ad una visione 'meccanicistica' della mente umana, anticipata dal funzionalismo omuncolare dei su citati Dennett e Lycan e recentemente ripresa dal neuro-funzionalismo di Jesse Prinz, con esplicito rinvio al neo-meccanicismo (Prinz 2012). I fenomeni mentali di livello personale – coscienti, attribuiti ad un soggetto unitario, descrivibili nei termini della psicologia di senso comune – sono considerati come fenomeni emergenti, consistenti in null'altro che nelle operazioni di una molteplicità di meccanismi neurali sub-personali, le cui parti costituiscono a loro volta meccanismi, giù per innumerevoli livelli di organizzazione (sistemi di neuroni, sinapsi, singole cellule, molecole, ecc.), fino a raggiungere le stesse componenti di base di cui è fatta la materia priva di mente – la materia 'fisica' nel senso più ordinario del termine. La mente, e l'organismo che la possiede, sono, in breve, livelli estremamente sofisticati di organizzazione della materia fisica. Ciò vale anche, va sottolineato, per l'attività mentale cosciente (il citato libro di Prinz contiene proprio una teoria meccanicistica della coscienza).

I filosofi neo-meccanicisti tendono ad evitare l'espressione 'macchina', perché troppo associata a macchine artificiali esclusivamente meccaniche, di tipo 'push-pull' (Machamer et al. 2000). Ma questa è solo una convenzione linguistica. Un qualsiasi insieme di meccanismi è, in senso perfettamente intelligibile, una macchina. La mente, e l'organismo che la possiede, sono dunque una macchina, 'soltanto' una macchina. La coscienza è una proprietà emergente della macchina.

Certamente è una proprietà che appartiene a macchine biologiche come noi. Macchine biologiche come noi hanno tutte le nostre capacità cognitive di alto livello, incluse emozioni e coscienza, per il semplice fatto che *sono* noi.

Ma non c'è ragione per escludere a priori che la proprietà della coscienza non possa appartenere anche a macchine artificiali opportunamente costruite – e, a maggior ragione, a cyborg.

5. Angosciati dalla macchina

Negli ultimi paragrafi ho ripercorso (semplificandoli brutalmente) modelli scientifici e argomenti filosofici molto elaborati e spesso molto tecnici, che sono difesi e discussi in modo spassionato. Per molti, però, questi argomenti e modelli hanno un eco emotivo: il disagio, se non l'angoscia, provocato dall'immagine di macchine artificiali che torna indietro confondendosi con la nostra immagine. Lo stesso disagio dell'assimilazione alla macchina che aleggia attorno alle strategie delle 'macchine mai!' e delle 'inimitabili macchine biologiche'. Lo stesso malessere che molti (non tutti) provano di fronte alla prospettiva meccanicistica, al rappresentare sé stessi – il proprio io, il soggetto – come un assemblaggio di meccanismi sub-personali.

È su questi sentimenti che mi voglio soffermare adesso, per concludere.

Il film *Blade Runner* è stato, nell'immaginario di più di una generazione, l'icona del gioco di specchi delle macchine, e del sotteso rovesciamento di un grande potere in una grande angoscia (rovesciamento ancora più inquietante nel romanzo di Philip K. Dick che ha ispirato il film, *Do Androids Dream of Electric Sheeps?*). Il potere è quello di costruire macchine artificiali capaci di replicare il corpo e soprattutto lo spirito umano, in modo così perfetto da confondersi con l'originale. L'angoscia non è tanto quella di perdere il controllo della creazione – la completa autonomia della creazione è, piuttosto, il suo massimo perfezionamento, e quindi la massima estensione del potere del creatore. L'angoscia è piuttosto quella di scoprire di non essere chi credevamo di essere: non essere l'originale, ma una replica assemblata, per quanto quasi perfetta; di essere, noi stessi, 'soltanto' una macchina.

Non serve a dissipare questo senso di angoscia che la macchina, come i replicanti di *Blade Runner*, sia una macchina davvero sofisticata e speciale, priva delle caratteristiche degradanti che associamo alle macchine più familiari – non è una 'cosa' priva di soggettività e coscienza, la sua azione non è costretta entro le linee di pochi processi stereotipati, ha pensieri complessi e sensazioni ed emozioni intense, ecc. Ma è pur sempre 'soltanto' una macchina – anche se la portata del 'soltanto' e di ciò che taglia via si fa sempre più umbratile, ineffabile. Soltanto materia in mezzo ad altra materia? Soltanto un nodo provvisorio nella catena delle cause? Soltanto un assemblaggio di pezzi 'replicanti' privi di una intrinseca, originaria, necessaria unità?

Per parte di noi, sono 'soltanto' che non hanno proprio senso. Non c'è nient'altro che sembra possibile, nient'altro che abbia senso desiderare, e nessun disagio, malessere, angoscia o nostalgia nel sentirsi relegati 'soltanto' a questo.

Per altra parte di noi, il 'soltanto' ha invece il senso di una perdita insopportabile. Non tanto direttamente la perdita del senso di sé, quanto piuttosto la perdita dell'idea che il senso di sé sia una prospettiva indiscutibile, incrollabile, necessaria. Per molti, senza questa idea, il senso di sé traballa.

Questa idea è quello che i filosofi hanno pomposamente chiamato 'metafisica del soggetto.' Dal che si vede come certa metafisica sia una cosa molto concreta e molto piccina: una pillolina contro i brutti sogni.

Riferimenti bibliografici

Barsalou L.W. 1999. 'Perceptual Symbol Systems', *Behavioral and Brain Sciences*, 22, 577-660.

Bechtel W. 2008. *Mental Mechanisms. Philosophical Perspectives on Cognitive Neuroscience*, Routledge.

Bechtel W., Abrahamsen A. 2005. 'Explanation: A Mechanistic Alternative', *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36, 2005, 421-441.

Bechtel W., Abrahamsen A., Graham G. 1998. 'The Life of Cognitive Science'. In: Bechtel W., Graham G. (eds), *A Companion to Cognitive Science*, Blackwell, 1-104.

Bermúdez J.L. 2020. *Cognitive Science: An Introduction to the Science of the Mind*, 3rd ed., Cambridge University Press.

Block N. 1995. 'On a Confusion about the Function of Consciousness', *Behavioral and Brain Sciences*, 18, 1995, 227-287.

Churchland P.M., Churchland P.S. 2000. 'Could a Machine Think?', *Scientific American*, January 1990, 32-37.

Craver C.F. 2007. *Explaining the Brain. Mechanisms and the Mosaic Unity of Neuroscience*, Oxford University Press.

Craver C.F., Tabery G., 'Mechanisms in Science', *Stanford Encyclopedia of Philosophy*, 2019, https://plato.stanford.edu/archives/sum2019/entries/science-mechanisms/

Dennett 1978. *Brainstorms. Philosophical Essays on Mind and Psychology*, Bradford Books.

Glennan S. 2017. *The New Mechanical Philosophy*, Oxford University Press.

Lycan W. 1981. 'Form, Function, and Feel', *Journal of Philosophy*, 78, 1, 1981, 24-50.

Miller G. 2003. 'The Cognitive Revolution. A Historical Perspective', *Trends in Cognitive Science*, 7, 3, 2003, 141-144.

Prinz J. 2012. *The Conscious Brain. How Attention Engenders Experience*, Oxford University Press.

Searle J.R. 1980. 'Minds, Brains, and Programs', *Behavioral and Brain Sciences*, 3, 3, 1980, 417-457.