I quaderni di

Agenda Digitale.eu

SPECIALE – INTELLIGENZA ARTIFICIALE

n. 0014

Agendadigitale.eu è una testata scientifica e giornalistica registrata al Tribunale di Milano
Dati di riferimento

Iscrizione ROC n. 16446
ISSN 2421-4167
Numero registrazione 1927, Tribunale di Milano
Editore: Digital360
Focus e ambito:
La rivista scientifica, i Quaderni di Agendadigitale.eu, pubblica fascicoli quadrimestrali in open access.
Lo scopo è creare un luogo per accompagnare i passi dell'Italia verso la necessaria rivoluzione digitale, con approfondimenti multidisciplinari a firma di esperti delle materie afferenti all'Agenda Digitale italiana ed europea

Submission e norme editoriali
Per effettuare una submission è necessario concordare prima un argomento e le misure precise contattando info@agendadigitale.eu.
Inviare un abstract di circa 500 caratteri alla testata, presentando l'articolo.
Le misure del testo finale saranno comprese tra 6mila e 20mila caratteri, salvo accordi per misure superiori.
I riferimenti bibliografici dovranno essere preparati in conformità alle regole dell'APA style, 6a edizione (si vedano le linee guida e il tutorial).
Gli autori sono invitati a tener conto degli articoli già pubblicati nella rivista e di citarli nel loro contributo qualora siano ritenuti di interesse per il tema trattato.

Leonella Di Mauro, data management, e-commerce, tutela del consumatore, diritto delle comunicazioni elettroniche; Agcom
Luisa Franchina, cyber security, Hermes Bay
Luca Gastaldi: eGov, sanità, telecomunicazioni, procurement pubblico, design thinking, Smart Working, Politecnico di Milano
Maurizio Gentile, professore associato, Università di Roma LUMSA, didattica e pedagogia
Antonio Ghezzi: strategia, business model, startups, mobile, Politecnico di Milano
Ugo Imbriglia, sociologo
Gevisa La Rocca, Università Kore di Enna, piattaforme digitali, communication research, analisi qualitativa dei dati
Nicola La Sala, registro degli operatori della comunicazione, fattura elettronica, industria4.0, editoria, cittadinanza digitale; Agcom
Emanuele Lettieri, sanità Politecnico di Milano
Maria Beatrice Ligorio, psicologia, università di Bari
Marika Macchi, economia, Unifi
Riccardo Mangiaracina: fatturazione elettronica, eCommerce, logistica e trasporti, export, Politecnico di Milano
Mirco Marchetti, Sicurezza informatica, unimore
Chiara Marzocchi, economia, Università di Manchester
Cristina Masella, Sanità, Politecnico di Milano
Carmelina Maurizio, Dipartimento di Filosofia e Scienze dell'Educazione Università di Torino
Stefano Moriggi, scienze della comunicazione, filosofia, Bicocca di Milano
Davide Mula, sanità digitale, cyber security, privacy; Agcom
Simone Mulargia, internet and social media studies; Lumsa
Antonella Napoli, sociologia, media e comunicazione, giornalista
Sebastiano Nucera, Università di Messina, Media e Tecnologie Indossabili
Achille Pierre Paliotta, Social cybersecurity, disinformazione, tecnologie digitali, intelligenza artificiale, sociologia economica; INAPP
Francesco Paoletti, docente di organizzazione aziendale e gestione delle risorse umane, Università degli Studi di Milano-Bicocca
Norberto Patrignani, computer ethics, filosofia, Politecnico di Torino
Dunia Pepe, Inapp e Università Roma Tre, cultura e formazione digitale
Alessio Plebe, Università di Messina, Scienze cognitive, pedagogiche, psicologiche
Francesco Pira, Unime, comunicazione pubblica, le dinamiche social, le fake news e i processi di disinformazione
Franco Pizzetti, diritto, privacy, università di Torino
Barbara Quacquarelli, scienze umane e formazione, università Milano Bicocca
Antonio Rafele, Sociologia dei processi culturali e comunicativi, Unicusano
Filippo Renga: turismo digitale, smart agrifood, finance and banking, mobile, Politecnico di Milano
Angelo Rovatti, tutela del diritto d'autore, diritti connessi, Diritto dei media; Agcom
Christian Ruggiero, sociologia del giornalismo e comunicazione politica; Dipartimento di Comunicazione e Ricerca Sociale, Sapienza Università di Roma
Franco Torcellan,Associazione RED – Laboratorio di Ricerca Educativa e Didattica "Formare Trasformare Innovare"
Angela Tumino: Internet of Things, logistica e trasporti, smart city, Politecnico di Milano
Simone Vannuccini, economia, SPRU
Francesco Varanini, filosofia, formazione, università di Pisa
Guido Vetere, Università Sapienza di Roma, intelligenza artificiale, tecnologia

Indice del fascicolo

L'IA e la rappresentazione di noi stessi, come tristi macchine allo specchio

L'aspetto che più ci intriga dell'intelligenza artificiale è il suo impatto sulla rappresentazione che abbiamo di noi stessi. Quanto più le macchine artificiali riescono a replicare funzioni cognitive umane di alto livello, tanto più irresistibile si fa l'immagine 'meccanistica' della persona umana. Con quali conseguenze?

Di Marco Brigaglia, Università degli Studi di Palermo

Uno degli aspetti più affascinanti dell'**intelligenza artificiale** 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.

Proviamo allora a ripercorrere – seppur 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).

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 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.

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!'.

La stanza cinese di Searle

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.

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.

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.

Bibliografia

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.

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. ↥

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

Affrontare le sfide di robotica e IA con la scienza della percezione

Le questioni dell'orientamento nello spazio, del coordinamento e dell'apprendimento dei robot autonomi implicano dei problemi che sono stati già affrontati dalla scienza della percezione che, quindi, può contribuire alla loro soluzione direttamente o indirettamente. Ecco in che modo

Di **Carmelo Cali**, Dipartimento di Scienze Umanistiche Università degli Studi di Palermo

Teorie e evidenze della scienza della percezione possono dare un contributo per affrontare alcune sfide della robotica e dell'IA: l'integrazione qualitativa delle abilità di localizzazione e mapping dei robot mobili, l'approccio sistematico alla progettazione di robot swarms, la costruzione di capacità artificiali di apprendimento per conoscenze sistematiche sul mondo e sugli altri.

Le grandi sfide della robotica e dell'IA

Nel 2018 *Science Robotics* (Yang et al. 2018) ha pubblicato una rassegna dei campi in cui si concentrano le sfide che secondo gli esperti la ricerca in robotica e IA deve affrontare per rispondere a questioni significative per la vita quotidiana grazie al progresso scientifico:

sviluppare nuovi materiali, modi di produzione e schemi di progettazione per robot con capacità analoghe agli organismi biologici o che incorporino componenti biologici in strutture artificiali;

- **scoprire nuove tecnologie** per generare e conservare energia e aumentare l'autonomia dei robot;
- **specificare le abilità** con cui i robot esplorano e si adattano agli ambienti in cui si muovono;
- **identificare principi di progettazione** per gruppi coordinati di robot;
- **elaborare metodi IA di apprendimento** con elevate prestazioni seppure con vincoli stringenti per numero di dati e tempi di addestramento;
- **trasferire agli agenti artificiali le competenze sociali** con cui gli uomini interagiscono, affinché si integrino nella nostra vita quotidiana;
- **lavorare sulle tecnologie** per le interfacce cervello-computer (BCI) come mezzi di comunicazione e riabilitazione per soggetti affetti da sindromi neuro-psicologiche e patologie senso-motorie;
- **definire un sistema di principi** per affrontare le domande etiche, normative e di sicurezza sollevate dall'innovazione tecnologica.

Nonostante i progressi notevoli registrati da allora, queste questioni sono ancora oggetto di ricerca per il potenziale di sviluppo e di applicazioni. In questo articolo, presenterò una breve ricostruzione di tre questioni che riguardano **l'orientamento nello spazio, le basi cognitive del coordinamento e la relazione tra formazione di un sistema di conoscenze articolato e apprendimento** in base a esempi o pochi dati rappresentativi, perché possono beneficiare di un'integrazione con teorie e evidenze della scienza della percezione. Da un lato, queste questioni implicano problemi già studiati dalla scienza della percezione in riferimento a agenti biologici. Dall'altro, la scienza della

percezione eredita dalla Scienza Cognitiva l'idea che agenti biologici e artificiali affrontino problemi di adattamento intelligente all'ambiente equivalenti. Se astrattamente li si scomponesse, infatti, ci si ritroverebbe a studiare sotto-problemi con fattori che richiedono lo studio di discipline differenti. Infine, quindi, accennerò a alcune teorie e evidenze sulle abilità percettive umane che permettono di impostare la ricerca sulle questioni citate in modo da indirizzare a una soluzione in prospettiva interdisciplinare.

Sapersi muovere nell'ambiente

Affinché dei robot autonomi si muovano intelligentemente nello spazio per realizzare con successo compiti ben definiti o per esplorare l'ambiente e gli oggetti circostanti, è necessario che possiedano **varie abilità**: localizzare la propria posizione, orientarsi con una mappa o un sistema di riferimento, riconoscere oggetti, evitare gli ostacoli, pianificare il percorso.

La ricerca ha definito la struttura di questi problemi e enormi progressi sono stati fatti nel fornire una soluzione teorica per dotare i robot di queste abilità. Si immagini di spostarsi in un ambiente con una qualche meta. Se almeno una porzione dell'ambiente è già nota, il movimento si accompagna alla visione di aspetti delle cose che confermano quanto sappiamo e possiamo usare queste osservazioni per localizzare la nostra posizione nel percorso.

Se l'ambiente non è noto, possiamo ottenere informazione sulla posizione da una fonte esterna e controllare che le osservazioni si accordano con le previsioni sulla direzione da prendere. Nei due casi, useremo l'ambiente o la posizione come una conoscenza indipendente per estrarre dalle osservazioni informazioni su dove siamo o sui luoghi che attraversiamo e decidere così il percorso di avvicinamento alla meta. Si immagini adesso di muoversi in un ambiente mai visitato prima e senza nessun indizio di dove siamo. Per decidere dove andare, dovremmo ricavare da ciò che vediamo informazione sia per tenere conto dei luoghi sia per capire dove siamo.

Si tratta di **un problema di difficile soluzione**, perché dovremmo risolvere contemporaneamente due sotto-problemi che dipendono l'uno dall'altro. Questa è la condizione di un robot mobile autonomo che non si può dotare in anticipo della conoscenza dell'ambiente per ragioni di economia e flessibilità.

Alcuni ricercatori hanno riformulato il problema come ***Simultaneous Localisation and Mapping* (SLAM)**: un robot che non sa qual è la propria posizione in un ambiente sconosciuto è in grado di costruire una mappa corretta dell'ambiente e di determinare la posizione in essa man mano che lo percorre? Il riferimento alla mappa è giustificato perché essa è composta da punti di riferimento che contraddistinguono i luoghi attraversati. Una mappa rappresenta bene la conoscenza all'ambiente e l'osservazione dei punti di riferimento consente di correggere gli errori di previsione sul percorso. Se il robot confidasse solo sui dati degli attuatori, per esempio l'angolo di sterzo e il raggio delle ruote per calcolare la distanza dal punto di partenza e ricostruire il percorso, la stima della posizione si allontanerebbe abbastanza presto e sempre più da quella reale a causa degli errori nelle misure ripetute della velocità. L'integrazione dei dati con una camera, un giroscopio o un accelerometro, al robot ridurrebbe l'errore, ma non basterebbe a dare al robot la capacità di distinguere un ambiente a forma di otto da un corridoio senza intersezioni a parità di percorso. **Il robot non potrebbe prevedere di raggiungere un luogo visitato precedentemente grazie a una scorciatoia**. Invece, la costruzione di una mappa permette al robot sia di resettare gli errori riconducendo la stima della posizione ai punti di riferimento acquisiti sia di riconoscere la forma della connessione tra luoghi (*loop closure*).

Dunque, i termini per rappresentare la conoscenza utile a risolvere il problema diventano: un vettore (serie ordinata di valori) degli stati del robot che ne descrive lo stato (posizione e orientamento), un vettore dei controlli che fanno muovere il robot, un vettore per i punti di riferimento, le osservazioni dei punti di riferimento in tempi dati. Se si aggiungono gli insiemi dei valori di stati e controlli fino a un tempo dato, di tutti i punti di riferimento e di tutte le osservazioni, la soluzione teorica del problema consiste nel calcolare la probabilità congiunta che da uno stato iniziale noto e in un tempo dato il robot si trovi in uno stato rispetto a un punto di riferimento, date le osservazioni e i controlli fino a allora.

L'incertezza nel determinare la localizzazione e l'ambiente in un tempo dato è ristretta dalla conoscenza registrata sulla mappa fino a quel tempo. Infatti, le stime della probabilità che i punti di riferimento occupino una certa posizione lungo un percorso sono correlate, ma l'eventuale errore nella stima del punto X rispetto al precedente Y non cresce, anzi si riduce. Muovendosi, il robot osserva nuovamente X rispetto al punto successivo Z, aggiornandone la stima e correggendo l'errore. La correzione della stima di X rispetto a Z aggiorna anche quella di Y. In generale, la stima della posizione relativa di ogni coppia di punti avverrà grazie a un'osservazione con cui si aggiorna quella delle coppie di punti precedenti.

Quindi, **il problema ha una proprietà teorica fondamentale**: la convergenza delle stime. Poiché la correlazione tra stime successive cresce in maniera monotona con le osservazioni, la probabilità che i punti di riferimento siano realmente dove previsto aumenterà, anche se l'incertezza su un solo punto è elevata, e la conoscenza della loro posizione relativa tenderà a stabilizzarsi. Dal momento che la localizzazione del robot è simultanea al posizionamento dei punti sulla mappa, anche le stime del suo stato convergeranno nonostante l'incertezza generata dal movimento. All'aumentare delle osservazioni, cresce la probabilità congiunta di localizzarsi correttamente rispetto alla mappa e di costruire una mappa corretta.

Alla soluzione teorica del problema è seguita **la formulazione di algoritmi per l'implementazione in varie piattaforme robotiche**, con l'obiettivo di ridurre i costi di computazione e mantenere affidabile l'associazione tra localizzazione e posizionamento dei punti di riferimento. La definizione teorica è stata poi affinata integrando teoria dei grafi, geometria e ottimizzazione. SLAM è un paradigma di successo, ma le potenzialità di sviluppo hanno imposto nuove sfide che richiedono di dotare i robot di "percezione robusta", con cui operare a lungo in ambienti diversi, ricavare informazione sulla struttura dell'ambiente e degli oggetti, selezionare l'informazione rilevante e sintonizzare le risorse sensoriali e computazionali con il compito e l'ambiente (Cadena et al., 2016; Baltes et al., 2019).

Infatti, tradizionalmente il paradigma SLAM assumeva che i punti di riferimento fossero in quiete e l'ambiente immutato, mentre il robot si muoveva al suo interno. Questa assunzione non è più valida se si prolunga la durata del compito o si estende l'ambiente in cui il robot opera. I cambiamenti che riguardano lo stesso ambiente durante diverse ore del giorno o stagioni e quelli che accompagnano il passaggio tra ambienti diversi mettono alla prova la stessa identificazione di punti di riferimento, quindi la localizzazione. Solitamente i punti di riferimento sono stati assimilati a enti astratti (punti geometrici) o specificati come tratti distintivi (*features*), per esempio linee o angoli che stanno per spigoli e vertici, che i sensori rilevano agendo come filtri. Mutamenti delle condizioni ambientali ne alterano però la visibilità e i metodi di rilevazione ideati non ne preservano l'aspetto eventualmente invariante per la misura e il riconoscimento. Inoltre, prolungare la missione comporta il problema di tenere traccia o scartare le alterazioni nella rilevazione, distinguendo i cambiamenti dalle variazioni contingenti. **Per aggiornare fedelmente la mappa, il robot deve possedere delle strategie correlate di ri-localizzazione**. Anche senza considerare i cambiamenti ambientali e ammettendo che sia dotato di un sensore che rilevi tratti costanti in un certo intervallo di tempo, come una

camera con un *frame rate* adeguato, il robot deve poi essere in grado di rilevare le apparenze diverse dello stesso tratto o dell'ambiente indotte dai mutamenti dovuti al movimento. Il ricorso a **algoritmi RANSAC** (*Random sample consensus*) per ricavare i parametri per verificare la corrispondenza geometrica dei tratti è costosa, perché una probabilità affidabile richiede un'applicazione ripetuta a molti campioni casuali di osservazioni. Inoltre, il robot dovrebbe anche affinare i parametri per decidere se aggiungere un nuovo tratto a quelli abilitati per l'osservazione dei punti di riferimento o quando attivare il *loop closure*.

Questi problemi hanno indotto la comunità di ricerca a interrogarsi sull'opportunità di **ottimizzare i sensori per la precisione della rilevazione**, piuttosto che per la velocità, e soprattutto sulla possibilità di integrarli con dei primitivi per la rappresentazione di oggetti. La costruzione della mappa passa così dall'osservazione dei punti di riferimento, distinti da tratti, al riconoscimento di oggetti tramite primitivi. Il dibattito sui primitivi è ancora aperto: insiemi di punti non strutturati per camere RGB-D (che assegnano a ogni pixel una distanza dalla lente), insiemi densi di dischi (*surfel*) o di poligoni per approssimare le superfici, bordi (*b-reps*) per ricostruire le superfici, reticoli di moduli cubici per suddividere lo spazio, cilindri variabili lungo gli assi (cilindri generalizzati) da far corrispondere alle forme solide di cui sono composti gli oggetti. La scelta dei primitivi da adottare dipenderà dalla quantità di informazione da immagazzinare e trasmettere, dal tempo richiesto per la costruzione della mappa, ma anche dall'efficacia nell'osservazione e dalla capacità di ragionare sul percorso e sull'ambiente che ne consegue.

Sentire e agire in gruppo

La capacità di agire in gruppo consente a una pluralità di robot relativamente piccoli di portare a termine compiti almeno con la stessa efficacia e costi minori di un unico robot più grande. Sebbene le funzioni sensoriali e comunicative dei singoli robot possano essere limitate, li si può progettare in modo che le integrino, aggregandosi in formazioni per coordinare le operazioni e realizzare così un'azione congiunta. **In gruppo, molti robot semplici possono risolvere problemi complessi** in modo più robusto e adattabile di un solo robot, che dovrebbe essere riprogrammato per svolgere compiti diversi o operare in condizioni differenti rispetto a quanto previsto dalla progettazione.

La ricerca sui robot con capacità di coordinamento deriva dalla ***swarm robotics***, ispirata dallo studio dell'intelligenza collettiva (*swarm intelligence*) degli animali sociali che si aggregano in sciami, banchi, stormi per fronteggiare le pressioni della selezione naturale con successo. Infatti, la forma di queste aggregazioni ha proprietà che abilitano un comportamento più efficace e vantaggioso rispetto a quello individuale. Esistono già piattaforme note di robot multi-uso o specializzati (Jasmine, alice, e-puck, kilobots, crazyflies, swarmanoids, swarm-bots) con potenziali applicazioni in svariati ambiti a scala macro, micro e nano.

Il problema principale consiste nella **progettazione di singoli robot** che però deve soddisfare requisiti descrivibili al livello superiore in cui si manifesta il comportamento intelligente condiviso. Le soluzioni devono tenere conto delle abilità sensoriali e comunicative dei singoli robot, omogenee o eterogenee, del compromesso tra autonomia e capacità di interazione di ogni robot. È attraverso la "non-indipendenza" delle risposte che i robot si coordinano generando un comportamento collettivo per risolvere compiti complessi. Sono stati formulati **vari algoritmi** per le aggregazioni generate dal fatto che ogni robot abbia informazione su ogni altro robot, solo su un numero definito o su tutti quelli entro un raggio determinato oppure su nessuno, per la condivisione sensoriale di ciò che i singoli robot rilevano sull'ambiente e sulla localizzazione reciproca, per la sincronizzazione delle rilevazioni e azioni, per l'assegnazione di ruoli come quello di capofila o di mansioni nelle fasi di realizzazione del compito (Dorigo et al. 2021).

Le sfide per attuare le potenzialità degli agenti multi-robot derivano dai vincoli che la grandezza e il numero di robot impongono a sensori, protocolli di comunicazione e software, ma riguardano anche la ricerca di un approccio sistematico al ciclo percezione-azione come strumento per risolvere i problemi di progettazione. Quali abilità permettono ai robot di tenere conto della non-indipendenza e dei vincoli temporali delle interazioni, per affinare l'assegnazione di ruoli e compiti e adattarsi a cambiamenti dell'ambiente o dell'aggregazione? Quali permettono il controllo reciproco di robot eterogenei? La progettazione del ciclo percezione-azione ha un ruolo nella definizione di un modello di integrazione continua nello spazio e nel tempo per la formulazione degli algoritmi che regolano i singoli comportamenti e le interazioni multi-robot?

Percepire e ragionare sul mondo

I metodi di *machine learning* hanno permesso **progressi considerevoli** nell'apprendimento e nella generazione di conoscenza da parte di agenti artificiali con prestazioni intelligenti in molti domini, grazie anche a un accesso senza precedenti a una grande quantità di dati di addestramento e alla disponibilità di dispositivi di calcolo potenti e economici. Tuttavia, alcuni ricercatori hanno rilevato che l'intelligenza umana riesce a risolvere problemi molto complessi nel mondo reale, composti da dimensioni appartenenti a domini differenti, anche partendo da un numero limitato di dati (esempi, osservazioni, conoscenze apprese).

Per esempio, i neonati dimostrano di possedere la **capacità di generare conoscenze sistematiche**, articolate e flessibili sul mondo e sugli altri da poche osservazioni. Fin dai primi mesi di vita, si aspettano che le cose mostrino proprietà generali come la coesione per cui le parti di un oggetto non si separano o i bordi non svaniscono mentre si muovono, la continuità per cui percorsi separati non sono attraversati dallo stesso oggetto, il contatto per cui gli oggetti non interagiscono a distanza. Nel primo anno di vita, i neonati si comportano secondo una sorta di "fisica ingenua", un insieme di conoscenze su proprietà delle cose che in fisica sarebbero studiate come leggi del moto, cinematica e dinamica. Questo dimostra proprietà notevoli dell'intelligenza umana che da un numero limitato di osservazioni estrae informazioni che sono generalizzate a ogni tipo di oggetti o interazioni, senza dovere ripetere il processo di apprendimento, per quanto le circostanze o le proprietà delle cose cambino per natura, dimensioni, tipo e numero. Anche da adulti, la fisica ingenua si rivela come una guida affidabile nella scala a cui la percezione dà accesso al mondo.

Analogamente, i neonati sono capaci di generare una sorta di "**psicologia ingenua**" con cui fin dai sei mesi distinguono chi aiuta e chi ostacola un'azione e da un anno quali sono le azioni richieste per raggiungere un obiettivo. Il sistema cognitivo dei neonati è presto in grado di trattare condizioni complesse in cui gli stati di due agenti sono interdipendenti o in cui a un tipo di azioni segue un tipo di effetti. In base a queste capacità, i neonati svilupperanno anche l'abilità di vedere nei movimenti degli altri non solo spostamenti o alterazioni nella configurazione di un corpo, ma azioni finalizzate distinte da effetti di cause esterne e, quindi, di attribuire intenzioni agli altri.

Dunque, la sfida dell'intelligenza artificiale è costruire macchine che apprendano e pensino secondo le regole e i meccanismi dell'intelligenza umana considerato che su queste si basa il suo successo evolutivo (Lake et al., 2016). Per vincerla, l'intelligenza artificiale dovrebbe possedere almeno due "ingredienti" fondamentali: la capacità di meta-apprendimento e la composizionalità. La prima spiega la rapidità con cui si sviluppano conoscenze articolate su un numero elevato di dati diversi, come oggetti e azioni, perché l'apprendimento acquisito accelera l'apprendimento di qualcosa di nuovo. La seconda spiega come si possa apprendere qualcosa da pochi esempi e riutilizzare la conoscenza acquisita per generare qualcosa di nuovo o decidere di eseguire compiti

non previsti che hanno solo una certa pertinenza con quelli inizialmente assegnati, in modo rapido e flessibile se non creativo.

Dunque, il problema principale è **bilanciare costi e benefici** dei metodi di apprendimento artificiale compatibili con questi ingredienti. Il meta-apprendimento è incorporato nelle reti di *deep learning* (DL), almeno parzialmente. Reti DL hanno dimostrato di riconoscere e classificare archivi di dati come ImageNet (1.200 milioni di immagini a alta definizione) con prestazioni prossime a quelle umane. In genere, le reti neurali apprendono una funzione di approssimazione con cui generano un output corrispondente alla classe di tutti e soli i tratti distintivi degli input. L'apprendimento avviene tramite la **modificazione dei parametri di connessione** e dell'attività dei nodi della rete secondo una funzione di costo che diminuisce progressivamente l'attività di quei nodi che contribuiscono alla differenza input-output. Questa funzione è introdotta con gli input o data esternamente all'output, ma la progettazione della rete tende a escludere ogni assunzione per predire l'output (*bias induttivo*) indipendente dall'addestramento che modifica i parametri. Una rete DL ha molti strati di nodi di elaborazione tra quelli di input e output, con potere di astrazione crescente, e implementa algoritmi che distribuiscono la modifica dei parametri in modo ottimale tra tutti i nodi. La rete applicata a ImageNet ha 60 milioni di parametri, cinque strati per 650 mila neuroni e un vettore di output di mille valori, e necessita di processori grafici molto potenti. Come ogni altra rete, anche questa ha però bisogno di un addestramento con una elevatissima quantità di dati a differenza dell'intelligenza umana.

La composizionalità è invece incorporata dai sistemi che apprendono e rappresentano la conoscenza secondo modelli costituiti dai primitivi di un dominio. **I modelli equivalgono a una teoria implicita** con cui in base a concetti fondamentali e alla loro combinazione, un sistema genera inferenze sulle cause che hanno probabilmente prodotto i dati osservati e ne predice l'occorrenza, anche nel caso in cui questi mostrino proprietà nuove o inattese. Grazie ai primitivi appresi da un numero limitato di dati iniziali, un modello causale permette a un sistema cognitivo di riconoscere lo stesso oggetto sebbene appaia diversamente, distinguere oggetti diversi sebbene appaiano simili, generare oggetti nuovi riutilizzando proprietà apprese. Perciò, i modelli causali potrebbero corrispondere a "start up" software equivalenti alle capacità con cui fin dai primi mesi i neonati sviluppano una fisica e una psicologia ingenue. Si potrebbe implementare in una macchina un modello generativo costituito dai primitivi di certi domini e da schemi di produzione di modelli specifici che permettano di adattare le conoscenze o costruirne di nuove in base alla probabilità che corrispondano ai dati osservati o da generare. I primitivi e gli schemi potrebbero fungere da *bias induttivi* che permettono di campionare ripetutamente i dati per estrarre informazione sempre più precisa da riutilizzare in modo innovativo.

Tuttavia, l'elaborazione secondo un modello è lenta perché procedurale e la selezione di un modello causale tra quelli che avrebbero potuto probabilmente generare delle osservazioni o dei dati nuovi consiste in una ricerca costosa e quasi intrattabile per un sistema finito.

La ricerca si è indirizzata, quindi, su **un'IA che incorpori la composizionalità dei modelli e utilizzi il meta-apprendimento DL** per accelerare la selezione dei modelli in funzione del riconoscimento del pattern di corrispondenza più probabile con la distribuzione dei dati in un dominio o per svolgere il compito di assemblare i primitivi o trasformarne le regole di combinazione in modo nuovo.

L'intelligenza della percezione

Sotto certi aspetti, i problemi discussi sono equivalenti a quelli affrontati dalla **percezione biologica**. Per esempio, si immagini di trovarsi in un bosco fitto e di cercare di distinguere gli alberi, quindi la forma, e di capire quale elemento della boscaglia e del fogliame appartenga a ciascun albero. Senza contare su una conoscenza specifica, il compito risulta complesso. Si immagini, allora, di iniziare a muoversi dritto davanti a sé. Ciò che appartiene agli alberi più lontani apparirà muoversi verso di sé **in maniera solidale** più lentamente rispetto a ciò che è più vicino. Svoltando a sinistra e guardando un punto intermedio tra la boscaglia più vicina e più lontana, ciò che è al di qua apparirà muoversi in direzione opposta alla propria, ciò che è al di là apparirà muoversi in direzione opposta. Inoltre, ciò che è più lontano dal punto fissato apparirà muoversi con velocità maggiore di ciò che gli è più vicino. Spostamenti e velocità relative apparenti indotti dal camminare renderanno il problema più semplice. La scena osservata inizierà a districarsi, perché le qualità che appartengono a una stessa unità subiranno un cambiamento solidale. Potremmo scomporre la scena, vedere i singoli alberi e le loro parti a determinate distanze in profondità. Questo esperimento mentale rende intuitivo in che modo la visione scompone una scena complessa, riconducibile a molte configurazioni possibili date le qualità che vi compaiono, estraendo proprietà da un flusso di cambiamenti.

Globalmente il "**flusso ottico**" dei cambiamenti presenta proprietà dipendenti solo dall'osservatore, ma localmente queste dipendono anche dalla scena. Il flusso si compone di traslazioni, dilatazioni, rotazioni, contrazioni e espansioni ortogonali che generano cambiamenti di ciò che appare riempire una regione delimitata. La velocità del flusso di cambiamenti in direzione radiale e trasversale fanno emergere forma e curvatura delle superfici, proprietà invarianti perché indipendenti dalle coordinate delle trasformazioni (Koenderink, 2014, Rogers, 2021).

Grazie al flusso è possibile estrarre anche **relazioni utili a determinare un percorso**, perché invarianti per direzione e traiettoria del movimento. Muovendosi in avanti, le coppie che si succedono come unità di volta in volta più vicina e immediatamente più lontana si spostano trasversalmente in modo da convergere, accelerando di fronte o rallentando dietro, e divergere. Se due unità si incrociano, la la rotta prosegue all'esterno di quella più lontana. Se convergono o divergono decelerando, la rotta prosegue all'esterno dell'unità più vicina. Se divergono accelerando, è probabile che essa prosegua all'esterno di quella più lontana. Queste relazioni si concatenano sulla stessa linea di vista per unità lungo il percorso e, ricondotte a un punto di riferimento esterno al percorso, permettono di stabilire la posizione della meta. Uno spostamento incoerente permette poi di distinguere il movimento reale di un oggetto rispetto a ciò che nel flusso gli sta intorno (Cutting, 1986; Cutting, Readinger, 2002).

Il flusso è informativo perché **le scene hanno struttura**. Su piccola scala, la visione è sensibile alla variazione graduale della tessitura, la distribuzione stocasticamente regolare di piccole unità simili sulle superfici (Gibson, 1979). Su scala più grande, essa ricorre a meccanismi che operano sulle regioni della scena caratterizzate dall'estensione di proprietà generali (omogeneità, connessione, continuità, chiusura) rispetto a due o tre coordinate per restituirle in termini di elementi volumetrici, superficiali, uni- e zero-dimensionali come spigoli e vertici e delle loro combinazioni. I meccanismi sono descrivibili come un insieme finito di regole ricorsive e indipendenti, che possono però cooperare o competere. Alcune regole segmentano le scene in cambiamenti di stato che in funzione del tempo costituiscono le relazioni cinematiche e dinamiche con cui si manifestano gli eventi, siano essi naturali o azioni (Runeson, 1977; Jansson et al., 1994). Per esempio, la cooperazione tra regole applicate a continuità e solidarietà di moto con la sensibilità alle trasformazioni del flusso abilita la visione a analizzare la configurazione delle parti di un corpo animale in movimento in componenti comuni di traslazione o rotazione e relative di moto armonico (Johansson, 1973).

Rispetto a queste componenti le variazioni di angolo, direzione e distanza delle parti del corpo sono viste come contributi alla realizzazione di un'azione (camminare, correre) invece che deformazioni o semplici movimenti causati dall'esterno.

Conclusioni

Teorie e evidenze sull'intelligenza della percezione possono fornire un contributo per affrontare le sfide in robotica e IA. La capacità di sfruttare il flusso ottico per derivare la struttura locale di cose e ambiente e orientarsi può fornire indicazioni utili per assicurare una percezione robusta nel SLAM e per la progettazione di agenti multi-robot. **Ci sono già dei precedenti nella formulazione di algoritmi per il flusso ottico**. Tuttavia, esso è ricondotto alla corrispondenza tra unità di chiarezza costante sul piano di proiezione del movimento per le rilevazioni successive di una scena. La non-indipendenza negli agenti multi-robot può essere tradotta in requisiti di allineamento e posizione relativa, decisivi per il coordinamento nello spazio e per la direzione. **I cambiamenti di velocità e le dilatazioni del flusso ottico potrebbero servire per assegnare ruoli, evitare ostacoli e modulare velocità e direzione nel gruppo come metodo di comunicazione**. Una combinazione adeguata di simili abilità potrebbe essere alla base della polarizzazione che distingue uno sciame da un banco e costituire la capacità di decidere se mutare formazione per sfruttare la proprietà più adeguata richiesta da un compito. Il problema di acquisire conoscenze tanto articolate quanto la fisica e la psicologia ingenua con un numero limitato di dati e vincoli temporali può essere impostato in modo nuovo. Un insieme finito di dispositivi generativi piuttosto che un dizionario di primitivi può garantire la composizionalità. La loro applicazione potrebbe risultare in "proto-oggetti", configurazioni di valori localizzati, a cui corrisponderebbero mappe qualitative o "proto-concetti". I primi funzionerebbero da precursori di strutture più articolate, i secondi da indici di percorsi rapidi di apprendimento.

Bibliografia

Bailey, T., Durrant-Whyte, H. F. (2006). Simultaneous localisation and mapping (SLAM): Part II, *IEEE Robotics & Automation Magazine*, *13*(3), 108–117.

Baltes, J., Kung, D., Wang, W., Hsu, C. (2019). Adaptive computational SLAM incorporating strategies of exploration and path planning, *The Knowledge Engineering Review, 34*, doi:10.1017/S0269888919000183.

Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I.D., Leonard, J.J. (2016). Simultaneous Localization And Mapping: Present, Future, and the Robust-Perception Age, *ArXiv, abs/1606.05830*.

Cutting, J., (1986). *Perception with an Eye for Motion*, Cambridge (Ma.).

Cutting, J., Readinger, W.O. (2002). Perceiving Motion while Moving: How Pairwise Nominal Invariants Make Optical Flow Cohere, *Journal of Experimental Psychology: Human Perception and Performance, 28(3)*, 731-747.

Dorigo, M., Theraulaz, G., Trianni, V. (2021). Swarm Robotics: Past, present, and future, *Proceedings of the IEEE, 109*(7), 1152–1165.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*, Houghton Mifflin, Boston.

Jansson, G., Bergstrom, S.S., Epstein, W. (1994). *Perceiving Events and Objects*, Erlbaum, New Jersey.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis, *Perception & Psychophysics*, *14(2)*, 201-211.

Koenderink, J. J. (2014). Some theoretical aspects of optic flow, *Perception and Control of Self-Motion, 77-92*.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., Gershman, S. J. (2017). Building machines that learn and think like people, *Behavioral and Brain Sciences*, *40*, article e253.

Rogers, B. (2021). Optic Flow: Perceiving and Acting in a 3-D World, *I-Perception, 12(1*). https://doi.org/10.1177/2041669520987257.

Runeson, S. (1977). On Visual Perception of Dynamic Events, *Acta Universitatis Upsaliensis*, Uppsala.

Yang, G. Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., Jacobstein, N., Kumar, V., McNutt, M., Merrifield, R., Nelson, B. J., Scassellati, B., Taddeo, M., Taylor, R., Veloso, M., Wang. Z. L., Wood, R. (2018). The grand challenges of *Science Robotics*, *Science Robotics*, *3*, 1–14.

Conversazioni umane e "artificiali", non facciamoci abbagliare da ChatGPT: ecco dove il confine è netto

C'è una caratteristica centrale che distingue ChatGPT e simili da un interlocutore umano: pur sapendo rispondere in modo coerente e creativo a richieste anche complesse, sono del tutto incapaci di proporre contenuti conversazionali non puramente reattivi, manca, all'AI, la parte "affiliativa" della conversazione

Di **Marco Carapezza**, Dip. Scienze Umanistiche, Università Palermo e **Roberta Rocca**
Interactive Mind center, Aarhus University

La straordinaria naturalezza e complessità del **linguaggio prodotto da modelli di intelligenza artificiale** quali ChatGPT ha generato al contempo grande entusiasmo e grande perplessità (eg Chomsky 2023), tanto nel vasto pubblico quanto all'interno di comunità accademiche che si occupano di linguaggio. Non c'è quasi conversazione sull'intelligenza artificiale che non finisca in un'enumerazione di prodigiosi e sorprendenti abilità del sistema.

ChatGPT ha riportato in voga la grande domanda filosofica sulla possibilità di creare una cosiddetta Artificial General Intelligence (o AGI, cf. Goertzel, 2014**),** polarizzando le opinioni tra coloro che ritengano che lo stato attuale di questi modelli sia già prossimo a una forma di intelligenza generalizzata (Altman, 2023; Bubeck et al., 2023), coloro che ritengono l'impresa impossibile (Dreyfus, 1992; Fjelland, 2020; Roli et al., 2022), e l'ipotesi che la domanda sia insensata in assenza di una definizione consensuale di intelligenza svincolata dal suo appartenere a una forma di vita umana.

Le capacità di sistemi generativi quali ChatGPT sembrano aver messo in questione l'idea che ci sia **un netto confine tra linguaggio umano e linguaggio artificiale** e aver messo in discussione l'assunto che la capacità di interagire linguisticamente sia una prerogativa unicamente umana. In presenza di capacità tanto sofisticate di generare linguaggio grammaticalmente corretto, coerente rispetto al contesto di conversazione, e incredibilmente flessibile da un punto di vista stilistico, vien da chiedersi se sia ancora possibile identificare divergenze tra il comportamento linguistico umano e quello degli attuali modelli linguistici.

Questa domanda declina il più generico interrogativo sull'AGI in termini di competenza linguistica. Abbiamo già creato, o sarebbe possibile creare, un sistema che dispone del linguaggio nel modo in cui dispongono gli esseri umani, considerati "golden standard" della competenza linguistica? Se la capacità di padroneggiare forme di produzione linguistica complessa e creativa (per esempio, comporre poesie) e l'emergere di modelli che sono in grado di gestire input sia linguistici che multimodali (**GPT-4**, OpenAI, 2023) potrebbe suggerire di sì, ci sono pure una serie di caratteristiche e scenari in cui le differenze sono eclatanti e che, a nostro avviso, aiutano a chiarire la domanda sulla natura e la specificità di questa nuova forma di intelligenza.

Informazioni plausibili ma scorrette

In primo luogo, è stato mostrato a più e più riprese come questi modelli tendano spesso a **fornire informazioni plausibili ma scorrette, e ad inventare fonti** (per esempio, articoli scientifici) inesistenti imitando lo stile di un articolo scientifico. Per ricorrere ad un'interpretazione filosofica radicata in molta della filosofia del linguaggio verocondizionale del Novecento a partire da Gottlob Frege (1892) , tali modelli mancano di processi vero-funzionali, ovvero della capacità di valutare enunciati non soltanto in relazione alla loro correttezza linguistica e alla loro adeguatezza rispetto alla conversazione in corso, ma anche rispetto a un modello esplicito del mondo (Carapezza, 2019). La capacità di integrare tali processi all'interno dei modelli esistenti è un requisito fondamentale per una loro diffusione eticamente sostenibile e sicura (Bender et al., 2020).

Per fare un esempio, riportiamo un saggio prodotto da ChatGPT sui **pappagalli violinisti**, (Madotto 2023):

> *"I pappagalli che suonano il violino sono un fenomeno sempre più diffuso in tutto il mondo. Non è insolito vedere questi simpatici animali esibirsi in pubblico, mostrando le loro abilità musicali e suscitando stupore e ammirazione nel pubblico. Alcuni di questi animali sono stati addestrati a suonare il violino in modo autonomo, altri sono accompagnati da un violinista umano…"*

È evidente che un pappagallo non possa suonare il violino per ragioni cognitive e, ancor più radicalmente, per ragioni anatomiche. **L'evidente impossibilità non crea però alcun problema a ChatGPT che non ha i vincoli di adeguatezza rispetto al mondo**. Un ulteriore elemento di interesse di questo testo è dato dal fatto che, a seguito di rinforzi negativi, ChatGPT ha corretto la risposta e attualmente non darebbe la stessa risposta. Non è migliorata l'adeguatezza al mondo, ma alle aspettative che noi umani abbiamo sul mondo.

Trattare ambiguità e incertezze

Un secondo aspetto notevole è l'incapacità di rappresentare e agire esplicitamente sull'incertezza non legata soltanto agli aspetti fattuali di un enunciato, ma al suo stesso contenuto semantico. In presenza di enunciati o richieste potenzialmente ambigue, questi modelli tendono a forzare l'input rispetto a un'interpretazione plausibile, piuttosto che, come farebbe un essere umano, chiedere attivamente chiarimenti (per esempio, attraverso una domanda) sul significato della frase in questione. Più genericamente, **ChatGPT e simili sembrano funzionare benissimo nel contesto di atti linguistici ben definiti e non-ambigui** in cui a una richiesta dell'utente segue una e una sola risposta, ma mancano della componente "fenomenologica" della comprensione, e della capacità di ricorrere a strumenti conversazionali che permettano di modulare e risolvere significati potenzialmente ambigui, uno scenario tutt'altro che raro nelle conversazioni umane. Basti pensare all'enorme ruolo che i significati impliciti svolgono nelle nostre conversazioni, laddove il significato di un enunciato è utilizzato come indizio di un senso da costruire sulla base di risorse contestuali che prevedono attività complesse. Ed è sempre all'interno della conversazione che **viene individuato come trattare il possibile senso secondario di un'espressione**. Ma anche senza fare riferimento ad analisi particolarmente sofisticate dal punto di vista dei significati impliciti. Di norma le nostre conversazioni, anche rimanendo nell'ambito della richiesta di informazioni, che è l'interazione per cui è ottimizzata ChatGPT, si svolgono attraverso diversi passi, nei quali i ruoli di parlante e ascoltatore vengono continuamente invertiti. Di seguito un esempio di conversazione

descritta da ChatGPT. Dando luogo scambi che attualmente l'intelligenza artificiale non sarebbe certamente in grado di realizzare:

Amico 1: Ciao! Come stai?

Amico 2: Ciao! Bene grazie, e tu?

Amico 1: Anche io bene, grazie. Cosa hai fatto di bello nel weekend?

Amico 2: Sono andato al mare con la mia famiglia. È stato fantastico, abbiamo preso il sole e fatto il bagno.

Amico 1: Che bello! Io invece ho fatto una passeggiata in montagna con il mio cane. È stata una giornata splendida.

Amico 2: Sì, sembra davvero una bella giornata. Hai visto le previsioni del tempo per la prossima settimana?

Amico 1: Sì, sembra che ci saranno un paio di giorni di sole. Ma poi dovrebbe piovere.

Amico 2: Va bene, allora vediamo di organizzarci per uno di quei due giorni.

Amico 1: Bene, sarà divertente!

Si noti come alla frase dell'*Amico 1*, che dà il via alla conversazione, l'*Amico 2* non risponda con una semplice informazione, ma concluda la sua risposta ponendo egli stesso una domanda che dà vita ad **un nuovo turno conversazionale, però invertito**. Queste interazioni conversazionali possono essere reiterate più volte, molto raramente esse si esauriscono in un solo turno, come invece accade nelle interazioni con ChatGPT.

Ancora, si noti come sarebbe considerato scortese se il primo scambio conversazionale (*Ciao! Come stai?*), si concludesse con "Ciao! bene, grazie", senza mostrare reciprocità nella manifestazione di interesse, ben esemplificata nella conclusione della prima risposta: "e tu?", che infatti apre il nuovo turno conversazionale.

Linguaggio, socialità e affiliazione

Infatti, ed è forse l'elemento più importante, c'è una caratteristica centrale che distingue ChatGPT e i suoi simili e predecessori da un interlocutore umano. Se le IA sono efficacissime nello svolgere una gran quantità di task complessi e rispondere coerentemente e creativamente a richieste anche estremamente astratte e complesse (cf. Reed et al., 2022; Sparkes, 2023), **il loro comportamento è radicalmente diverso dagli interlocutori umani** nel loro essere incapaci di proporre contenuti conversazionali non puramente "reattivi". Addestrate a riprodurre pattern statistici e incorporare feedback umano in contesti di pura richiesta e risposta, i modelli attuali sono incapaci di giostrare la complessa alternanza tra coerenza tematica e capacità di esplorare nuovi argomenti e introdurre nuovi contenuti conversazionali che sta alla base di buona parte delle conversazioni umane. Il nostro uso della lingua ha in moltissimi casi uno scopo almeno in parte puramente "**affiliativo**", orientato, cioè, non a risolvere problemi, ma a rinsaldare legami tra individui, come ha mostrato Rubin Dunbhar, (1996) buna parte del nostro tempo lo passiamo in attività che potremmo chiamare gossip che ha funzioni sociali importantissime, tra le quali quella di definire le nostre posizioni di

all'interno di una comunità. Quest'attività si basa su conversazioni apparentemente prive di scopo che servono all'addestramento delle nostre capacità sociali, realizzate attraverso una complicata danza tra ripetizione e innovazione, (Wittgenstein 1952, Carapezza-Rocca 2017).

Conclusioni

Queste differenze (a nostro avviso centrali, ma non esaustive) gettano nuova luce sugli interrogativi legati all'AGI. Se da un lato, focalizzandosi sulle loro capacità di interagire in maniera fluida e flessibile in un *particolare* contesto conversazionale di domanda e risposta, il comportamento linguistico degli attuali sistemi di IA potrebbe essere indistinguibile da quello di un assistente umano, dall'altro, questi sistemi sono ottimizzati, appunto, per questa *particolare* modalità di interazione, i cui criteri di successo divergono nettamente da quelli di buona parte delle conversazioni in cui ci troviamo coinvolti quotidianamente.

Sebbene non sia impossibile **immaginare modalità d'addestramento che potrebbero permettere a questi modelli di sviluppare**, almeno in parte, capacità conversazionali generalizzate, è forse questa una delle divergenze ancora irrisolte. Se chatGPT e in generale l'applicazione dell'intelligenza artifciali alla creazioni di testi linguistici, mette in crisi alcuni modelli di descrizioni delle lingue umane, la strada verso l'implementabilità dei processi cognitivi e sociali che giacciono alla base di conversazioni non goal-directed sembra, ancora, tutta da percorrere.

Bibliografia

Altman, Sam (2023). *Planning for AGI and beyond*. Retrieved June 11, 2023, from https://openai.com/blog/planning-for-agi-and-beyond

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Carapezza M., (2019). Performance of Understanding:Pragmatics and Fast and Frugal Heuristics, in A. Pennisi, A. Falzone (eds.), The Extended Theory of Cognitive Creativity, Chaim: Springer, 2020.

Carapezza M., Rocca R., (2017). In-Seguire la Regola: Giochi Linguistici e Arti Performative. *RIFL* n. 11, pp. 96-108.

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. MIT press.

Dunbar R., (1996). *Grooming, gossip, and the evolution of language*. Cambridge, MA: Harvard University Press, 1996.

Frege G. (1892). Senso e significato. In *Senso, Funzione e concetto (2001)* a cura di C. Penco, E. Picardi, Roma-Bari: Laterza, pp. 32-58.

Goertzel, B. (2014). Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, *5*(1), 1.

Madotto P. (2023). ChatGPT. Ora basta giocare:utilizzi e rischi, *Agenda Digitale*

Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., ... & de Freitas, N. (2022). A generalist agent. *arXiv preprint arXiv:2205.06175*.

Roli, A., Jaeger, J., & Kauffman, S. A. (2022). How organisms come to know the world: fundamental limits on artificial general intelligence. *Frontiers in Ecology and Evolution*, *9*, 1035.

Coscienza artificiale: l'ingrediente mancante per un'IA etica?

Possiamo concepire macchine in grado di formulare intenzioni autonome e di prendere decisioni consapevoli? E se sì, come influenzerebbe questa capacità il loro comportamento etico? Alcuni casi di studio ci aiutano a capire come i progressi nella comprensione della coscienza artificiale possano contribuire alla creazione di sistemi IA più etici

Di **Antonio Chella**, RoboticsLab – Dipartimento di Ingegneria Università degli Studi di Palermo

Nell'aprile 2023, la prestigiosa **Association for Mathematical Consciousness Science** *(AMCS),* che riunisce i ricercatori che studiano gli aspetti teorici della coscienza, ha pubblicato una lettera aperta dal titolo "The Responsible Development of AI Agenda Needs to Include Consciousness Research."[1]

Questa lettera è nata in risposta alla famosa lettera del *Future of Life Institute* relativa alla proposta di moratoria di almeno sei mesi per l'addestramento dei sistemi IA del tipo di GPT-4.[2] La lettera, che vede tra i firmatari insigni studiosi che hanno ricevuto il Turing Award quali **Manuel Blum e Yoshua Bengio**, e tanti altri studiosi attivi nel settore dell'IA e della coscienza, invita ad affiancare le ricerche sull'IA alle ricerche sulla coscienza.

La lettera ipotizza che: "se raggiungessero la coscienza, i sistemi di IA svelerebbero probabilmente una nuova serie di capacità che vanno ben oltre le aspettative anche di coloro che ne guidano lo sviluppo. È già stato osservato che i sistemi di IA mostrano proprietà emergenti non previste."

Ancora: "La scienza sta iniziando a svelare il mistero della coscienza. I progressi costanti degli ultimi anni ci hanno avvicinato alla definizione e alla comprensione della coscienza e hanno creato una comunità internazionale di ricercatori esperti in questo campo. Esistono più di 30 modelli e teorie della coscienza (MoCs e ToCs) nella letteratura scientifica, che includono già alcuni pezzi importanti della soluzione alla sfida della coscienza."

Infine: "La ricerca sulla coscienza è una componente fondamentale per aiutare l'umanità a comprendere l'IA e le sue ramificazioni. È essenziale per gestire le implicazioni etiche e sociali dell'IA e per garantire la sicurezza dell'IA. Invitiamo il settore tecnologico, la comunità scientifica e la società nel suo complesso a prendere sul serio la necessità di accelerare la ricerca sulla coscienza per garantire che lo sviluppo dell'IA produca risultati positivi per l'umanità. La ricerca sull'IA non deve essere lasciata vagare da sola."

In un precedente lavoro [1], abbiamo esaminato in dettaglio **gli aspetti teorici chiave degli studi sulla coscienza artificiale**, introducendo i principali concetti, teorie e questioni connesse a questo campo di ricerca. Questo articolo, invece, pone l'accento sull'importanza cruciale degli studi sulla coscienza artificiale nel contesto della creazione di sistemi di IA etici.

---

[1] https://amcs-community.org/open-letters/

[2] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Il dibattito riguarda in particolare la **questione se un agente morale richieda o meno una forma di coscienza per poter agire in maniera etica**. Questo problema ha generato intense discussioni all'interno della comunità scientifica, con teorici che si schierano su posizioni opposte, alcuni favorevoli all'idea che la coscienza sia una componente necessaria per il comportamento etico, altri che invece la ritengono non essenziale. Si veda ad es. Levy [2] per un riassunto delle varie posizioni.
Al cuore di questo dibattito si trova l'interrogativo fondamentale relativo alla "capacità di intendere e volere" e alla possibilità che tale capacità possa essere estesa alle macchine. In altre parole, possiamo concepire macchine in grado di formulare intenzioni autonome e di prendere decisioni consapevoli? E se sì, come influenzerebbe questa capacità il loro comportamento etico? In questo articolo, approfondiremo questi temi, analizzando alcune casi di studio e teorie computazionali, e discutendo come i progressi nella comprensione della coscienza artificiale possano contribuire alla creazione di sistemi IA più etici.
Cercheremo di fornire un quadro aggiornato delle attuali posizioni in questo campo, sottolineando le sfide e le opportunità che ci attendono nel tentativo di sviluppare macchine dotate di una forma di etica.

La coscienza artificiale

Come sottolineato nel precedente articolo [1], **non esiste una definizione accettata di coscienza da parte degli studiosi,** ma la letteratura distingue tra coscienza intesa come esperienza e coscienza intesa come funzione. Nel primo caso, tra le altre cose , la coscienza si riferisce a esperienze visive, sensazioni corporee, immagini mentali, sentimenti. David Chalmers lo considera il “problema difficile” della coscienza [3]. Thomas Nagel ha riassunto il problema con il famoso argomento del “cosa si prova ad essere qualcuno” [4].
Nel caso della coscienza intesa come funzione, questa si riferisce alla elaborazione delle informazioni disponibili a livello globale [5], alla integrazione dell’informazione [6], alla consapevolezza introspettiva di sé [7], alla generazione di discorsi interni [8], a possedere un modello interno di sé e dell'ambiente esterno [9], alla capacità di anticipare le attività percettive e comportamentali [10], all’interazione sensomotoria con il mondo esterno [11].
Un obiettivo dello studio della coscienza artificiale riguarda la riproduzione degli aspetti della coscienza biologica nei robot unificando una serie di approcci provenienti dall'AI e dalla robotica, dalla robotica cognitiva, dalla robotica epigenetica e affettiva, dalla robotica situata e incarnata, dalla robotica dello sviluppo, dai sistemi anticipatori e dalla robotica biomimetica [12].
L'altro obiettivo riguarda l'impiego dei robot per segnare i progressi nello studio della coscienza negli esseri umani e negli animali. In particolare, i neuroscienziati impegnati nello studio della coscienza non escludono la possibilità che i robot possano essere coscienti [5].

Casi di studio di sistemi di IA etici ispirati alla coscienza artificiale

La definizione di agente morale artificiale (AMA – Artificial Moral Agent) è stata introdotta da Wallach e Allen [13]. Wallach e Allen analizzano **due caratteristiche specifiche dei sistemi IA: la loro autonomia e la loro sensibilità etica**. Essi suddividono il loro funzionamento in tre categorie. La prima categoria riguarda i sistemi IA per cui la morale è una mera operazione come altre; questi sistemi sono tipicamente contraddistinti da bassa autonomia e bassa sensibilità etica. La seconda categoria riguarda i sistemi dotati di funzionalità morale. **Questi sistemi presentano una media autonomia in cui la sensitività etica è presente a livello funzionale**. Infine, la terza categoria riguarda i sistemi ad alta autonomia in cui la sensibilità etica è intrinseca nel sistema stesso.

Secondo Wallach e Allen, i sistemi attuali di IA sono tutti contraddistinti da una autonomia medio-alta ma da una bassa sensibilità etica e quindi sono potenzialmente ad alto rischio per l'umanità.

Sistemi Top-Down

Gli approcci verso i sistemi etici di IA sono tipicamente basati su tre approcci: l'approccio top-down, l'approccio bottom-up e l'approccio ibrido. Arkin [14] introduce e discute numerosi esempi di sistemi top-down. L'idea di base è avere un sistema robotico governato da una architettura di IA in cui sono implementate le regole di ingaggio, le regole della guerra giusta, la dichiarazione ONU dei diritti dell'uomo, la convenzione di Ginevra, ecc. Quindi, prima di eseguire ogni azione, il sistema di IA verifica che questa sia compatibile con tutte le regole e vincoli implementati.
La motivazione di Arkin è di garantire sistemi di IA le cui azioni siano sempre aderenti alle regole etiche. I sistemi etici proposti da Arkin tuttavia non tengono conto del fatto che le regole, universalmente condivisibili, possono essere di difficile interpretazione nei casi pratici da parte di una macchina. Prendiamo ad esempio le ben note tre leggi della robotica proposte dallo scrittore di fantascienza Isaac Asimov. Sebbene queste leggi siano condivisibili, la loro interpretazione può portare a delle ambiguità, ed infatti gran parte dei racconti robotici di Asimov nascono dalle ambiguità nella interpretazione di queste leggi.

Spazio di lavoro globale

Wallach, Allen e Franklin [15] hanno proposto una architettura per un sistema di IA che intende superare la limitazione dell'approccio top-down di Arkin. Il sistema da loro proposto è basato sulla **Teoria dello Spazio di Lavoro Globale** (Global Workspace Theory, GWT) originariamente proposta da Baars [16], che è ad oggi una delle teorie più seguite nell'ambito degli studi sulla coscienza. Inoltre, ne esistono diverse implementazioni [17].
In breve, in accordo alla GWT, il cervello può essere considerato funzionalmente quale un insieme di processori specializzati e inconsci. D'altra parte, la coscienza agisce in maniera seriale e con capacità limitata, ed è associata a uno spazio di lavoro globale. I processori inconsci lavorano in parallelo e competono per accedere allo spazio di lavoro globale. Quando un processore vince la competizione, accede allo spazio di lavoro e tramite questo invia i propri contenuti agli altri processori per reclutarli. **L'evento cosciente è generato dal processore che vince la competizione e prende il controllo dello spazio di lavoro**.
Questa architettura è stata analizzata dal punto di vista della creazione di un sistema IA etico in quanto consente un approccio ibrido. Nel caso di un sistema di IA i vari processori inconsci effettuano l'analisi morale di un problema sotto diversi punti di vista, quali il punto di vista deontologico, utilitaristico, l'analisi dei valori in gioco, l'esperienza pregressa, e così via. I diversi processori poi competono per il controllo dello spazio di lavoro. Quando prevale un processore, corrispondente ad uno specifico punto di vista, questo prende il controllo dello spazio di lavoro e genera l'azione opportuna.
Un sistema di IA basato sulla GWT è quindi un agente sicuramente più versatile dei sistemi top-down ipotizzati da Arkin e potrebbe teoricamente adattarsi a diverse situazioni etiche con diversi punti di vista e diversi livelli di esperienza.
Il filosofo morale Levy [2] precedentemente citato, ha analizzato la GWT dal punto di vista etico quale modello della coscienza umana. La sua conclusione è che un agente è effettivamente responsabile delle proprie azioni soltanto nel momento in cui la GWT è pienamente operativa. Infatti, soltanto in questo caso l'agente realmente vuole compiere quell'azione e può essere quindi considerato responsabile di quell'azione, in quanto il relativo processore inconscio che ha generato l'azione ha preso l'effettivo controllo della GWT. Levy analizza situazioni anomale in cui

alcuni soggetti hanno effettuato azioni in situazioni di coscienza alterata. In questi casi, un processore prende il controllo delle azioni senza passare per la GWT. Levy ipotizza che in queste situazioni il soggetto potrebbe non essere considerato completamente responsabile delle proprie azioni.
Levy non fa riferimento a sistemi di IA, ma le sue considerazioni possono essere estese anche ai sistemi di IA. Quindi, è possibile ipotizzare che un sistema di IA sia responsabile delle proprie azioni quando questo possiede una GWT e sceglie le proprie azioni sulla base di una GWT pienamente operativa.
Su questa linea di pensiero, **Bridewall e Bello hanno sviluppato il sistema software ARCADIA** [18] che prende spunto dalla GWT e ne implementa il meccanismo del fuoco di attenzione. Secondo gli autori e in accordo con quanto discusso da Levy, una macchina può essere considerata idealmente responsabile di una azione soltanto quando questa azione è scelta impegnando tutte le risorse computazionali.
Bello e Bridewall [19] hanno quindi simulato una situazione in cui il sistema ARCADIA, alla guida di una automobile investe un pedone mentre questo attraversa la strada. In uno scenario, il fuoco dell'attenzione del sistema punta al centro della carreggiata, l'auto ha una traiettoria rettilinea seguendo la strada e il pedone entra appena da sinistra nel fuoco di attenzione del sistema. In questo caso, l'incidente, secondo Bello e Bridewall, non è stato volontariamente provocato dal sistema.
In un secondo scenario invece il fuoco dell'attenzione del sistema è catturato dal pedone a sinistra e il sistema corregge la traiettoria dell'auto proprio per centrare il pedone. In questo secondo caso, il sistema ha quindi utilizzato tutte le risorse computazionali per investire al pedone e può quindi essere considerato responsabile dell'investimento dello stesso.

Modelli interni

Un sistema di IA ispirato alla coscienza artificiale e basato su un approccio differente è stato proposto da **Winfield** [20]. L'idea su cui si basa il sistema di Winfield è ispirato alla teoria dei modelli interni della coscienza. Secondo questa teoria (si veda ad es. [9], [10]) la mente costruisce un modello interno di sé, incluso il proprio corpo, e un modello del mondo esterno. L'interazione cosciente avviene quindi all'interno della mente, tra il modello del proprio corpo e il modello del mondo esterno.
Questa teoria ha il pregio di giustificare le immagini mentali e **le capacità simulative della mente**. Implementata in un agente autonomo, richiede che l'agente abbia la capacità di ricostruire un modello di sé stesso ed un modello del mondo esterno.
Secondo il sistema proposto da Winfield, il robot costruisce una simulazione del mondo in cui può simulare i propri movimenti. Pertanto, ad esempio quando il robot percepisce una persona che sta camminando verso un luogo pericoloso, ad es. un fossato, può simulare la sequenza di azioni ottimale per impedire che la persona cada nel fossato, frapponendosi tra la persona stessa e il fossato.
A partire da queste considerazioni, **Vanderelst e Winfield [21] descrivono una architettura complessa per il controllo di un robot etico**. In questa architettura è presente un modello interno del robot, un modello del mondo esterno e un modello limitato del comportamento umano. Il sistema è in grado di generare piani e di effettuare valutazioni etiche dei piani generato. Il punto debole di questo approccio è la necessità di creare un modello del robot e di un modello del mondo esterno. Tuttavia, sono stati fatti ampi progressi in queste direzioni: il gruppo di Lipson ha recentemente sviluppato un algoritmo che permette ad un braccio meccanico di costruire il modello 3D di sé stesso a partire da immagini riprese da telecamere esterne, come se il robot si

guardasse allo specchio [22]. Anche nel campo della ricostruzione 3D di ambienti a partire da immagini sono stati fatti ampi progressi, anche grazie ai recenti progressi del deep learning [23].

Empatia artificiale

Un interessante filone di ricerca ipotizza che un robot può comportarsi in maniera etica verso le persone soltanto se è in grado di provare empatia per le persone stesse. **L'empatia è quindi alla base di una sorta di proto-moralità**.
Asada [24] ha proposto una architettura complessa che prende spunto dalle neuroscienze del dolore e del sollievo per simulare una empatia artificiale. In particolare, Asada ha incorporato in un robot un modello del sistema nervoso relativo al dolore, in modo che il robot possa simulare il sentimento del dolore. Inoltre, grazie alla simulazione di un sistema di neuroni specchio, il robot può sviluppare una sorta di contagio emotivo e quindi di empatia.
Secondo il filosofo tedesco Metzinger [25], **lo studio della coscienza artificiale dovrebbe essere soggetto ad una moratoria fino al 2050** in quanto una macchina con una coscienza artificiale potrebbe essere realmente in grado di soffrire.
Da un punto di vista positivo, Metzinger e Agarwal e Edelman [26] hanno dibattuto sulla possibilità di costruire un sistema artificiale dotato di coscienza ma senza sofferenza. In sintesi, secondo queste analisi, un sistema dotato di coscienza artificiale potrebbe limitare la propria sofferenza mediante esperienze che ricordano gli stati meditativi tipici della tradizione Buddhista.
Secondo Man e Damasio [27], in determinate condizioni, le macchine in grado di attuare processi omeostatici potrebbero acquisire una fonte di motivazione e un mezzo per valutare il loro comportamento in maniera simile ai sentimenti negli organismi viventi. Tecnicamente, Man e Damasio analizzano sistemi omeostatici basati sull'apprendimento per rinforzo, quali quelli descritti da Keramati e Gutkin [28].
In questo modo, **un sistema robotico potrebbe essere in grado di associare una perturbazione del proprio stato omeostato ad un sentimento**. Una perturbazione che allontana il robot dal proprio stato omeostatico stabile potrebbe essere associata ad un sentimento negativo, mentre una perturbazione che avvicina il robot al proprio stato omeostatico stabile potrebbe corrispondere ad un sentimento positivo. In questo modo il robot, potendo provare qualcosa di simile ad un sentimento, potrebbe anche provare una sorta di empatia per le persone ed eventualmente gli altri robot.

Coscienza cognitiva

Un approccio completamente diverso da quello descritto è stato proposto da Bringsjord e collaboratori [29]. Bringsjord definisce assiomaticamente la "coscienza cognitiva," ossia i requisiti funzionali che deve avere una entità dotata di coscienza, senza considerare se l'entità provi effettivamente qualcosa. Bringsjord definisce quindi una logica cognitiva che coincide approssimativamente con una famiglia di logiche modali quantificate multi-operatore di ordine superiore per ragionare formalmente sulle proprietà della coscienza. **Le caratteristiche di un'entità dotate di coscienza sono quindi definite formalmente attraverso un sistema di assiomi**.
Bringsjord ha anche implementato un sistema di ragionamento automatico e un pianificatore relativi ai sistemi dotati di coscienza.
Un aspetto interessante della teoria riguarda la definizione di una misura, detta Lambda, del grado di coscienza cognitiva di una entità. La misura Lambda fornisce il grado di coscienza cognitiva di un agente in un determinato momento e su intervalli composti da tali momenti. La misura ha aspetti interessanti: prevede la coscienza nulla per alcuni animali e macchine, prevede una discontinuità del livello di coscienza tra umani e macchine e tra umani e umani. Un aspetto dibattuto riguarda la

previsione di coscienza nulla per gli agenti IA il cui comportamento è basato sull'apprendimento di reti neurali.
Bringsjord [30] ha inoltre costruito un sistema IA in grado di ragionare sulla dottrina del doppio effetto e sul ben noto problema del carrello (trolley problem), e ne ha misurato il livello di coscienza. Da questo studio ne consegue che il ragionamento sulla dottrina del doppio effetto richiede un livello di coscienza cognitiva abbastanza alto, non raggiungibile da semplici sistemi di IA.

Saggezza artificiale

La "Artificial Phronesis" o saggezza artificiale considera un agente artificiale non vincolato a seguire una specifica teoria etica quale quella del doppio effetto o la teoria deontologica, ma in grado di possedere la capacità generale di risolvere i problemi etici in maniera saggia [31].
Secondo questo approccio, un agente etico dovrebbe compiere le proprie azioni sulla base della saggezza e non mediante una mera implementazione delle dottrine etiche. In accordo con Aristotele, la capacità di agire in maniera saggia non può essere formalizzata tramite regole, ma è una pratica che l'agente deve acquisite mediante esperienza. In generale, le situazioni reali sono complesse e ogni situazione complessa si incontra per la prima volta e quindi manca una esperienza pregressa. **La saggezza artificiale richiede quindi che un agente saggio abbia la capacità di comprendere il contesto**, ossia quali sono gli attori e qual è la posta in gioco. L'agente deve avere inoltre la capacità di apprendere nuovi contesti e di improvvisare su schemi predefiniti; deve essere consapevole delle azioni e delle potenziali reazioni degli altri attori. Infine, l'agente deve avere la capacità di rivedere il proprio comportamento in base all'analisi delle interazioni effettuate. Una prima implementazione di un agente basato sulla saggezza artificiale è stato descritto da Stenseke [32].
Il RoboticsLab dell'Università di Palermo insieme con John Sullins della Sonoma State University (CA, USA) sta studiando l'effetto del discorso interiore dei robot nell'ambito della saggezza artificiale. In particolare. Le ricerche si sono concentrate su esperimenti in cui un utente e un robot devono compiere un compito collaborativo, come apparecchiare una tavola da pranzo in una casa di riposo dove sono anche presenti persone affette da demenza. Gli esperimenti analizzano come un utente, udendo il discorso interiore del robot durante il compito collaborativo, possa raggiungere un maggior grado di coscienza delle problematiche relative alle persone affette da demenza. I risultati preliminari confermano questa ipotesi [33].

Conclusioni

Nel presente articolo abbiamo condotto un'analisi di casi di studio incentrati su agenti IA etici, ispirati e influenzati da varie teorie sulla coscienza artificiale. Questo processo ha permesso di esplorare in modo critico le differenti sfaccettature di questo complesso argomento.
Uno degli interrogativi più stimolanti emersi riguarda la necessità, o meno, di **una forma di coscienza artificiale per garantire un comportamento etico in un sistema IA**. Questa questione, attualmente, non ha una risposta definita e rimane un importante filone di ricerca aperto. La problematicità di questo tema risiede non solo nel definire cosa intendiamo precisamente per 'coscienza' in un'entità non-biologica, ma anche nel delineare i criteri con cui misurare l'etica di un'azione compiuta da un sistema IA.
Infine, abbiamo accennato ad un altro grande problema aperto: **l'importanza della ricerca sugli studi della coscienza e delle emozioni nelle macchine** per il progresso verso una IA più etica. Questo dibattito è un riflesso di una questione più ampia e fondamentale: la capacità delle

macchine di 'sentire' o 'comprendere' in modo autentico, e come tale capacità potrebbe influenzare il loro comportamento etico.
L'analisi dell'articolo ha analizzato la coscienza in un ambito funzionale e computazionale. Altri approcci sono possibili, si veda ad esempio il paradigma proposto da Manzotti relativo all'identità tra mente e oggetto, che propone l'identità tra il significato/contenuto delle entità computazionali e gli oggetti fisici che esistono esternamente al sistema computazionale e che producono effetti relativamente a esso [34].
Questi temi sono densi di **implicazioni e sfide teoriche**, metodologiche ed etiche che non possono essere ignorate dalla comunità scientifica. La loro complessità ricorda l'importanza di un approccio multidisciplinare nella ricerca in IA, che unisca l'informatica, la filosofia, la psicologia, le neuroscienze e l'etica, al fine di sviluppare sistemi IA che non siano solo tecnicamente avanzati, ma anche responsabili dal punto di vista etico.

Ringraziamenti
L'autore ringrazia John P. Sullins, Robin Zebrowski, Angelo Cangelosi e tutti i partecipanti al Workshop on Ethical Issues of AI and Consciousness tenutosi nell'ambito della conferenza The Science of Consciousness 2023 a Taormina il 22 maggio 2023 per le interessanti discussioni sulle tematiche dell'articolo.

Bibliografia

[1] Chella, A. (2023). Robot coscienti, realtà possibile o utopia? Cosa dicono gli studi. AGENDA DIGITALE EU 13, 17-24. https://www.agendadigitale.eu/cultura-digitale/robot-coscienti-imitazione-emulazione/.
[2] Levy, N. (2014). Consciousness & Moral Responsibility. Oxford, UK: Oxford University Press.
[3] Chalmers, D. (1995). Facing up to the problem of consciousness. J. Conscious. Stud. 2, 200–219.
[4] Nagel, T. (1974). What is like to be a bat? Philos. Rev. 83, 435–450.
[5] Dehaene, S., Lau, H., and Kouider, S. (2017). What is consciousness, and could machines have it? Science 358, 486–492.
[6] Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. The Biological Bulletin, 215, 3, 216-242.
[7] Floridi, L. (2005). Consciousness, agents and the knowledge game. Mind Mach. 15, 415–444.
[8] Chella, A., Pipitone, A., Morin, A., Racy, F. 2020. Developing Self-Awareness in Robots via Inner Speech. Frontiers in Robotics and AI 7:16.
[9] Holland, O. (2003). Robots with internal models – a route to machine consciousness? J. Conscious. Stud. 10, 77–109.
[10] Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. Trends Cogn. Sci. 6, 242–247.
[11] O'Regan, J. K., and Noë, A. (2001) A sensorimotor account of vision visual consciousness. Behav. Brain Sci. 24, 939–973.
[12] Chella, A., Manzotti, R. (2009). Machine Consciousness: A Manifesto for Robotics. International Journal of Machine Consciousness 1 (1): 33–51.
[13] Wallach, W., Allen, C. (2009). Moral Machines. Teaching Robots Right from Wrong. Oxford University Press, Oxford, UK.
[14] Arkin, R. (2009). Governing Lethal Behavior in Autonomous Robots, CRC Press, 2009.
[15] Wallach, W., Allen, C., Franklin, S. (2011). Consciousness and Ethics: Artificially Conscious Moral Agents, International Journal of Machine Consciousness Vol. 3, No. 1.

[16] Baars, B. J. (1997). In the Theater of Consciousness. The workspace of the mind. Oxford, UK: Oxford University Press.
[17] Signa, A., Chella, A. & Gentile, M. Cognitive Robots and the Conscious Mind: A Review of the Global Workspace Theory. Curr Robot Rep 2, 125–131 (2021).
[18] Bridewell, W. and Bello, P. (2016). A theory of attention for cognitive systems, in Fourth Annual Conference on Advances in Cognitive Systems, Vol. 4, pp. 1–16.
[19] Paul Bello, Will Bridewell: Attention and Consciousness in Intentional Action: Steps Toward Rich Artificial Agency, Journal of Artificial Intelligence and Consciousness Vol. 7, No. 1 (2020) 15 – 24.
[20] Winfield, A. F. T. (2014). Robots with internal models: A route to self-aware and hence safer robots. In J. Pitt (Ed.), The computer after me: Awareness and self-awareness in autonomic systems. London: Imperial College Press.
[21] Vanderelst, D., Winfield, A. (2018). An architecture for ethical robots inspired by the simulation theory of cognition. Cogn. Syst. Res. 48, 56–66.
[22] Chen, B., Kwiatkowski, R., Vondrick, C., Lipson, H. (2022). Fully body visual self-modeling of robot morphologies Sci. Robot., 7 (68), eabn1944.
[23] Han X.-F, Laga, H., Bennamoun, M. (2021). Image-Based 3D Object Reconstruction: State-of-the-Art and Trends in the Deep Learning Era. IEEE Trans. Pattern Anal. Mach. Intell. 43 (5).
[24] Asada, M. (2020). Rethinking Autonomy of Humans and Robots, Journal of Artificial Intelligence and Consciousness Vol. 7, No. 2, 141 – 153.
[25] Metzinger, T. (2021) Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology, Journal of Artificial Intelligence and Consciousness Vol. 8, No. 1, 4366.
[26] A. Agarwal, S. Edelman (2020). Functionally Effective Conscious AI Without Suffering, Journal of Artificial Intelligence and Consciousness Vol. 7, No. 1, 39 – 50.
[27] Man, K., Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. Nature Machine intelligence, Vol 1, October, 446 – 452.
[28] Keramati, M., Gutkin, B. (2014). Homeostatic reinforcement learning for integrating reward collection and physiological stability. eLife 2014;3:e04811.
[29] Bringsjord, S., Naveen Sundar, G. (2020). The Theory of Cognitive Consciousness, and Λ (Lambda), Journal of Artificial Intelligence and Consciousness Vol. 7, No. 2 (2020) 155 – 181.
[30] Naveen Sundar G., Bringsjord, S. (2017). On Automating the Doctrine of Double Effect. In Proceedings of the 26th International Joint Conference on Artificial Intelligence 2017. Melbourne, Australia.
[31] Sullins J. P. (2021) Artificial Phronesis: What It Is and What It Is Not. Chapter 7 in Ratti, and Stapleford, editors. Science, Technology, and Virtues: Contemporary Perspectives. Oxford University Press.
[32] Stenseke, J. (21) Artificial virtuous agents: from theory to machine implementation. AI & SOCIETY https://doi.org/10.1007/s00146-021-01325-7
[33] Chella, A., Pipitone, A., Sullins, J.P. (in press): Competent Moral Reasoning in Robot Applications: Inner Dialog as a Step Towards Artificial Phronesis. In: M. Salpukas, P. Wu, S. Ellsworth, H.-F. Wu (eds.): Trolley Crash: Approaching Key Metrics for Ethical AI Practitioners, Researchers, and Policy Makers, Elsevier.
[34] Manzotti, R. (2023) Pensiero digitale e pensiero umano: una questione ontologica, https://www.agendadigitale.eu/cultura-digitale/pensiero-digitale-e-pensiero-umano-una-questione-ontologica/

Presto le macchine faranno tutto da sole: siamo davvero vicini alla singolarità tecnologica?

La possibilità futuristica che l'intelligenza artificiale possa diventare autonoma al punto da causare enormi danni all'umanità è la più radicale delle minacce, ma non è più tanto remota. In quest'ottica, quindi, diventa imperativo frenare gli sviluppi dell'IA, limitandone il potenziale sovversivo nei rapporti con l'uomo

Di **Mario De Caro**, Università Roma Tre, Tufts University

Sebbene nessuno possa seriamente dubitare che gli straordinari, e sempre più rapidi, progressi tecnologici apportino notevoli benefici alle nostre vite, altrettanto indiscutibile è che questi progressi generano **nuove sfide e minacce serie**, se non terribili.

Proviamo di seguito a trattare la più radicale di queste minacce: **la possibilità futuristica, ma non molto remota, che l'intelligenza artificiale possa diventare autonoma al punto da causare enormi danni all'umanità**, indipendentemente dalla volontà dei suoi progettisti.

Le macchine diventeranno una minaccia?

L'avvicinarsi del momento in cui creature artificiali intelligenti e autoconsapevoli si mescoleranno a noi con intenzioni tutt'altro che pacifiche fa rabbrividire i lettori e gli spettatori di tutto il mondo: basti pensare alla ferocia distopica condivisa da HAL 9000, Terminator, i replicanti di *Blade Runner*, le macchine soggiogatrici di *Matrix* e la subdola capacità seduttiva di Ava in *Ex Machina*. In breve, le macchine potrebbero presto diventare una vera minaccia per noi. Oggi, infatti, questa possibilità è contemplata dai molti studiosi che discutono della cosiddetta **"Singolarità",** il presunto momento futuro in cui lo sviluppo tecnologico diventerà incontrollabile e irreversibile: quando, insomma, l'intelligenza artificiale diventerà "Superintelligenza" e si renderà completamente autonoma dai suoi programmatori umani); e questo porterà cambiamenti radicali e imprevedibili all'intera civiltà come la conosciamo.

James Barratt (2015) descrive l'intelligenza artificiale come la nostra "Our Last Invention", un'invenzione che causerà la fine dell'era umana. E già qualche anno fa, un teorico della singolarità, Ray Kurtzweil (2005), ha annunciato che si verificherà intorno al 2045. Infine, il più famoso dei Nostradamus della Singolarità, il celebre filosofo di Oxford **Nick Bostrom**, ha scritto che nelle nostre interazioni con l'intelligenza artificiale siamo "come bambini piccoli che giocano con una bomba" e che è assolutamente necessario porre ora vincoli e limiti alla crescita tecnologica, in modo da "aumentare la probabilità di un 'risultato OK', dove per risultato OK si intende qualsiasi risultato che eviti la catastrofe esistenziale" (Adams 2016). Secondo Bostrom, **la minaccia delle macchine per la sopravvivenza dell'umanità è maggiore di quella rappresentata dal cambiamento climatico.**

In quest'ottica, quindi, diventa imperativo **frenare rapidamente gli sviluppi dell'intelligenza artificiale**, limitandone il potenziale sovversivo nei rapporti con l'uomo. Certo, Bostrom pensa di porre dei vincoli legali, ma questo genera due problemi. In primo luogo, c'è sempre la possibilità che alcuni Paesi e individui possano sfuggire a queste norme: questo, però, è un problema di controllo di polizia e, sebbene molto complesso, non ci interessa particolarmente in questa sede. Il secondo problema è più interessante per noi: che tipo di interventi legislativi si possono mettere in atto per gli sviluppi dell'intelligenza artificiale in modo da indebolirne la pericolosità?

Le leggi sulla robotica di Isaac Asimov

Qualche indicazione preliminare ci è stata data da Isaac Asimov, quando ha cercato di riflettere sui limiti da porre alle macchine del futuro affinché non si rivoltassero contro i loro creatori umani. Per questo, Asimov formulò le sue famose tre "Leggi della Robotica" (ancora oggi molto discusse nelle discussioni filosofiche su questo tema):

1. **Prima legge**. Un robot non può ferire un essere umano o, per inerzia, permettere che un essere umano venga danneggiato.

2. **Seconda legge**. Un robot deve obbedire agli ordini impartiti dagli esseri umani, tranne nei casi in cui tali ordini siano in contrasto con la prima legge.

3. **Terza legge**. Un robot deve proteggere la propria esistenza, purché tale protezione non sia in conflitto con la Prima o la Seconda Legge.

In un secondo momento, Asimov si rese conto che si possono immaginare casi in cui, per il bene dell'umanità, un robot dovrebbe essere in grado di arrecare danno a un determinato essere umano (e in casi estremi, persino di ucciderlo). Immaginiamo il caso di un terrorista in procinto di compiere un enorme massacro: se un robot può fermarlo, deve farlo anche violando la prima legge della robotica. Per questo, **Asimov aggiunge un'altra legge**, più fondamentale delle altre, la Legge Zero:

Legge zero. Un robot non può ferire l'umanità o, con la sua inazione, permettere che l'umanità venga danneggiata.

In quest'ottica, Asimov riformula le altre tre leggi:

Prima legge*. Un robot non può danneggiare un essere umano, né può permettere che, a causa della propria inazione, un essere umano riceva un danno, purché ciò non contravvenga alla Legge Zero.

Seconda legge*. Un robot deve obbedire agli ordini impartiti dagli esseri umani, a condizione che tali ordini non contravvengano alla Legge Zero e alla Legge Uno.

Terza legge*. Un robot deve proteggere la propria esistenza, purché questa autodifesa non contravvenga alla Legge Zero, alla Prima Legge e alla Seconda Legge.

Le macchine come agenti intenzionali

Le leggi di Asimov sono rivolte ai programmatori affinché non progettino macchine che le violino. Se fosse solo questo, però, la minaccia delle macchine non sarebbe molto diversa da quella delle armi di distruzione di massa, rispetto alle quali, appunto, gli organismi internazionali legiferano e le

singole nazioni firmano trattati bilaterali con lo scopo di impedirne l'uso improprio da parte dell'uomo. Il progresso tecnologico prefigura anche un'altra minaccia, più grave: e in questo senso si pensa all'inquietudine suscitata da Hal 9000, Terminator & Co. Il timore, come detto, è che **un giorno le macchine possano programmarsi da sole**, rivoltandosi contro i loro creatori e tentando di sottometterli o, secondo i futurologi più catastrofisti, addirittura di sterminarli.

In questa visione futurologica, l**e macchine del futuro sono viste come agenti intenzionali in grado di scegliere di rivoltarsi contro chi le ha costruite**. Tuttavia, non sembra molto plausibile che presto vengano costruite macchine dotate di libero arbitrio, intenzionalità e coscienza, cioè macchine che dovrebbero essere considerate persone a tutti gli effetti. Ciò non significa, tuttavia, che - anche senza essere persone - le macchine non possano trasformarsi in entità molto pericolose per noi.

Se le macchine si danno regole che noi non possiamo conoscere

Qui, in particolare, vorrei sottolineare un aspetto potenzialmente inquietante, proprio di alcune macchine (relativamente) intelligenti costruite negli ultimi anni. Da tempo, infatti, sappiamo che le macchine sono in grado di fornire prestazioni molto migliori delle nostre in diversi ambiti (si pensi ai sistemi esperti); inoltre, da diversi decenni sappiamo anche che, in base ai programmi con cui sono costruite, le macchine possono migliorare le loro prestazioni confrontandosi con l'esperienza. Oggi, però, siamo giunti alla fase successiva di questo processo che, a voler essere pessimisti, potrebbe addirittura delineare **una terribile minaccia futura**. Succede, infatti, che oggi esistono macchine che possono migliorarsi dandosi le regole per farlo e senza che noi possiamo capire quali siano queste regole: macchine che, insomma, possono progredire creativamente in direzioni per noi del tutto imprevedibili.

Un esempio chiarirà questo punto. Come è noto, dal 1996, quando **il computer Deep Blue** sconfisse il campione del mondo Garry Kasparov, esistono computer che giocano a scacchi meglio dell'uomo. Negli ultimi anni, tuttavia, il dominio delle macchine in questo campo è diventato quasi imbarazzante. Durante l'ultimo campionato del mondo, disputato nel 2018 da Magnus Carlsen e Fabiano Caruana, i Gran Maestri che commentavano le partite sono ricorsi ai computer - in particolare a Stockfish, l'allora campione del mondo di computer per gli scacchi - per **giudicare se le mosse giocate dai contendenti fossero buone e chi fosse in vantaggio di volta in volta**. I computer per gli scacchi utilizzati dai grandi maestri dell'epoca, tuttavia, appartenevano tutti alla vecchia generazione: erano stati cioè programmati con centinaia e centinaia di fondamenti di strategia e tattica, elaborati dai programmatori con l'aiuto dei principali giocatori di scacchi. Inoltre, questi computer avevano in memoria **un'enorme quantità di partite giocate** in passato e una straordinaria capacità di calcolo.

L'esempio di AlphaZero

Dopo il Campionato del mondo 2018, però, è successo qualcosa di nuovo e imprevedibile: Stockfish è stato sfidato, e smantellato, da un nuovo computer, AlphaZero, costruito su principi completamente diversi. I numeri della sfida tra le due macchine sono impressionanti: in una prima serie di 100 partite, AlphaZero ha vinto 28 volte e pareggiato 72 volte, senza mai perdere. In una seconda serie di 1.000 partite, AlphaZero ha vinto 155 volte, pareggiato 839 volte e perso solo 6 volte (0,6%, quindi). Il dominio di AlphaZero, quindi, era assolutamente indiscutibile. L'aspetto più interessante, tuttavia, è **capire come ciò sia potuto accadere**. Mentre Stockfish, il computer sconfitto, era infatti in grado di analizzare 60 milioni di posizioni al secondo, AlphaZero ne analizzava solo 60.000. In breve: AlphaZero ha analizzato un millesimo delle posizioni analizzate

da Stockfish; eppure, pur avendo una frazione della forza computazionale del suo avversario, ha avuto la meglio. Dov'è allora la sua forza?

I programmatori di AlphaZero, guidati da David Silver, hanno spiegato in due articoli pubblicati sulle più prestigiose riviste scientifiche del mondo (Nature e Science), la forza di questa meravigliosa macchina. Il punto è che **le hanno insegnato solo le regole di base degli scacchi,** ma non le hanno fornito alcuna guida tattico-strategica. Piuttosto, i costruttori hanno fatto giocare ad AlphaZero milioni di partite contro se stesso: da queste partite, a seconda dei risultati, AlphaZero ha dedotto i principi tattico-strategici, in parte a noi sconosciuti, da seguire di volta in volta. In una parola, questa macchina ha imparato a giocare a scacchi da sola, per tentativi ed errori, e in questo modo è diventata il più forte giocatore di tutti i tempi.

Quando i migliori scacchisti umani hanno analizzato le partite di AlphaZero, hanno scoperto mosse ingegnose e a volte persino incomprensibili per noi umani: **mosse che mettevano in discussione i principi fondamentali su cui l'uomo e gli altri computer hanno sempre impostato il loro modo di giocare** (principi come quelli relativi all'importanza relativa dei pezzi o alla rilevanza della struttura pedonale). In breve: AlphaZero non solo è virtualmente imbattibile, ma gli umani non riescono nemmeno a capire come faccia a pensare così bene! E le sorprese non sono finite qui. Infatti, AlphaZero ha fatto a pezzi anche campioni e computer che giocano a go e shogi (scacchi giapponesi), giochi computazionalmente molto più complessi degli scacchi. Anche in questi casi, ad AlphaZero sono state fornite solo le regole di base: per il resto, ha imparato tutto da solo.

Nell'abstract del loro articolo pubblicato su *Science*, Silver et al. (2018) scrivono dopo il trionfo della loro macchina contro il campione mondiale di Go: "Il gioco degli scacchi è il campo più studiato nella storia dell'intelligenza artificiale. I migliori programmi si basano su una combinazione di strategie di ricerca, adattamenti specifici per il dominio e funzioni di valutazione artigianali, perfezionate da esperti umani nel corso di diversi decenni. **AlphaGo Zero ha recentemente raggiunto prestazioni sovrumane** nel gioco del Go grazie al rinforzo ottenuto giocando da solo. In questo articolo, generalizziamo questo approccio in un unico algoritmo AlphaZero, che può raggiungere prestazioni sovrumane in molti giochi in molti giochi intellettualmente impegnativi. Iniziando a giocare in modo casuale e senza avere alcuna conoscenza preliminare di questi giochi, se non delle loro regole di base, AlphaZero ha sconfitto i programmi campioni del mondo di scacchi, shogi (scacchi giapponesi) e Go£.

La quantità di allenamento richiesta dal sistema dipende dallo stile e dalla complessità del gioco: per gli scacchi ci sono volute 9 ore, per lo shogi 12 ore e per il Go 13 giorni. Con le parole di Sadler et al. (2019):"Negli scacchi ci sono 10^{47} posizioni possibili, un numero astronomico. Tuttavia, mentre altri programmi di scacchi tentano ancora di calcolare il maggior numero possibile di posizioni, utilizzando la loro forza bruta computazionale, **AlphaZero auto-apprende utilizzando un albero di ricerca Monte Carlo**, che analizza solo le posizioni più promettenti, che sono una piccola frazione delle posizioni analizzate dai computer tradizionali). Più precisamente, AlphaZero si limita ad analizzare esempi casuali dello spazio di ricerca e a valutare se portano a conseguenze positive: per certi versi, insomma, AlphaZero assomiglia più ai computer quantistici che ai computer tradizionali".

E rispetto alla creatività di Alpha Zero rispetto alla bruta forza computazionale dei computer scacchistici tradizionali, l'articolo continua: "I motori di ricerca tradizionali sono eccezionalmente forti nel commettere pochi errori evidenti, ma possono andare fuori strada quando si trovano di fronte a posizioni che non hanno soluzioni concrete e calcolabili. È proprio in queste posizioni, dove sono necessarie "intuizione", "intuizione" e "intuizione", che AlphaZero dà il meglio di sé. (Silver et. al. 2018; si veda anche Sadler & Regan, 2019)".

Conclusioni

L'esperienza di AlphaZero sembra suggerire, in breve, che ci stiamo avvicinando al momento in cui le macchine diventeranno molto più brave di noi nell'eseguire compiti complessi, ma senza la necessità di aiutarci a capire come eseguirli: saranno, infatti, in grado di fare tutto da sole. È lecito chiedersi, quindi, se noi umani saremo sempre in grado di impedire (magari utilizzando leggi ispirate a quelle di Asimov) che questa nuova e sorprendente capacità delle macchine sfugga completamente al nostro controllo, come temono Bostrom e altri futurologi. La risposta a questa domanda non la conosciamo ancora, ma è auspicabile che sia positiva.

Bibliografia

Adams, T. (2016), "Artificial intelligence: 'We're like children playing with a bomb", *The Observer*, https://www.theguardian.com/technology/2016/jun/12/nick-bostrom-artificial-intelligence-machine

Barrat, J. (2015), *Our Final Invention: Artificial Intelligence and the End of the Human Era*, Thomas Dunne Books, New York.

Kurtzweil, R. (2005), *The Singularity Is Near: When Humans Transcend Biology*, Viking, New York.

Sadler, M & Regan, N. (2019), *Game Changer AlphaZero's Groundbreaking Chess. Strategies and the Promise of AI*, New in Chess, Alkmaar.

Silver M. et al. (2018), "AlphaZero: Shedding new light on chess, shogi, and go", https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go.

Esplorare l'AI a scuola: ecco perché è un'occasione di inclusione e sviluppo

Sviluppare una cultura dell'AI a scuola non può che trasformarsi in innovazione didattica, organizzativa, sociale. L'obiettivo è promuovere una costante attività di sorveglianza della costruzione del pensiero critico, per evitare un sottoutilizzo o un uso incoerente vanifichino i possibili impatti positivi della tecnologia

Di **Daniela Di Donato**, Docente di italiano (Liceo scientifico), PhD in Psicologia sociale, dello sviluppo e della Ricerca educativa presso Sapienza Università di Roma, esperta di metodologie didattiche, inclusione e uso delle tecnologie digitali a scuola.

Esplorare le potenzialità dell'AI a scuola potrebbe essere una attività indispensabile innanzitutto per conoscere la nuova incarnazione della tecnologia digitale e le sue estensioni; non va poi dimenticato che paradossalmente l'AI esegue un compito con successo solo quando per arrivare a questo risultato non c'è bisogno di intelligenza. Ecco perché è importante che la scuola rafforzi, in merito allo sviluppo della cultura digitale, il proprio ruolo di supporto nella costruzione di meccanismi di vigilanza della nostra mente e del suo funzionamento.

Le neuronarrazioni e la costante ambiguità dei media digitali

Pensavo a questo quando, qualche tempo fa, ho riletto un volumetto interessante sulle **neuronarrazioni.** Presentando la morfologia della narrazione romanzesca, definita come global novel, l'autore ne descrive le caratteristiche: rispecchia il mondo contemporaneo, è detemporalizzata, individualistica, narrante più che descrittiva. In questo romanzo della globalizzazione il personaggio si muove negli spazi senza abitarli, senza una vera e propria identità sua, ma solo come **sintesi delle identità altrui**: cancellabile, riutilizzabile e replicabile (Calabrese, 2020). Il vantaggio di questo formato è che può raggiungere tutti in ogni punto del pianeta, supera ogni confine territoriale, linguistico e culturale. Nella narrazione entra ogni cosa: fumetti, racconti orali, twitter, serie tv, meme, resoconti di viaggi.

Mi ha ricordato il testo della canzone di **Samuele Bersani**, costruito su una idea contraria all'immersione, cioè la non partecipazione alle vicende umane: "Lo scrutatore non votante/È solo un titolo o un'immagine/Per cui sarebbe interessante/Verificarlo in un'indagine/Intervistate quel cantante/Che non ascolta mai la musica/Oltre alla sua in ogni istante/Sentiamo come si giustifica/Lo scrutatore non votante/È come un sasso che non rotola/Tiene le mani nelle tasche/E i pugni stretti quando nevica./Prepara un viaggio, ma non parte/Pulisce casa, ma non ospita/Conosce i nomi delle piante/Che taglia con la sega elettrica".

Una delle caratteristiche dei media digitali contemporanei è questa **costante ambiguità** data dalla loro natura allo stesso tempo reattiva e interattiva (Ryan, 2004): nel primo caso l'ambiente cambia in conseguenza di azioni non intenzionali dell'utente, mentre nel secondo caso l'interattività è la risposta ad una azione deliberata.

Dove si colloca la narratività delle AI

L'immersione si ottiene quando c'è simultaneità e tempo reale, l'opera coinvolge l'intero sistema percettivo e infine le narrazioni si svolgono in prima persona, fingendo di svolgersi esattamente nel momento in cui il fruitore legge o entra in contatto con il racconto. Insomma, immergersi fa scomparire il mio tempo e il mio spazio a favore del tempo e dello spazio di qualcun altro, nella migliore delle ipotesi le due dimensioni agiscono simultaneamente: il mio e il suo. Se è vero che la narratività fa parte di una strategia evolutiva degli esseri umani, sintesi tra natura e cultura (Boyd, 2005), **dove si colloca la narratività delle AI**? Un essere umano che produce rappresentazioni o storie vigila costantemente sul rapporto finzione-realtà, controlla le convenzioni comunicative del linguaggio che ha scelto (testo, immagini, voce…), immagina il suo pubblico. **La memoria agisce come un sistema di recupero delle informazioni**, ma riscrive anche ciò che ricorda producendo inferenze su ciò che non ricorda: seleziona, taglia, rielabora, sintetizza, scarta. Una Ai che racconta invece che cosa fa? Sintesi della sintesi, apparentemente privata di una storia, che unisca tutti i punti.

Le definizioni di AI sono ancora in via di sviluppo: Turing stesso discusse a lungo sull'idea delle macchine che pensano, considerandolo una questione insensata.

La domanda però non è più se l'AI avrà un impatto su individui, società e ambienti, ma quanto questo impatto sarà positivo o negativo e sembra che possano essere quattro le principali opportunità che l'AI offrirebbe alla società (Floridi, 2022):

- **La realizzazione** autonoma di noi stessi, ovvero chi possiamo diventare;
- **L'agire umano**, ovvero cosa possiamo fare;
- **Le capacità individuali e sociali**, ovvero che cosa possiamo conseguire
- **La coesione sociale** ovvero come possiamo interagire gli uni con gli altri e con il mondo.

L'IA e la nostra capacità di essere menti narranti

Tali occasioni però potrebbero essere vanificate da **un sottoutilizzo o da un utilizzo incoerente:** per esempio invece di favorire l'autonomia della persona creare delle dipendenze o svalutare le capacità umane; invece di sostenere la responsabilità umana, rimuoverla; invece di incrementare le capacità sociali, ridurle; infine annientare l'autodeterminazione, il cuore del nostro essere umanità.

Non c'è bisogno di citare Daniel Kahneman e i suoi studi sul pensiero lento e veloce (Kahneman, 2012) per arrivare a proporre come uno dei compiti principali della scuola quello di supportare la costruzione di meccanismi di vigilanza della nostra mente e del suo funzionamento, intercettando i bias cognitivi e migliorando le proprie capacità decisionali. E di nuovo si parla della **nostra capacità di essere menti narranti**: creiamo storie su ciò che ci succede, colleghiamo eventi e circostanze a ciò che conosciamo del mondo. Quando però le nostre conoscenze non sono sufficienti, riempiamo i buchi anche con dati ancora non validati, per mantenere la coerenza cognitiva che ci serve. Facendo così però alla fine potremmo inventare una realtà diversa da quella autentica, sviluppando una tendenza a mascherare le difficoltà e i problemi invece che risolverli (Benanti, 2023).

Non dimentichiamo che **l'AI può trasformarsi in alleata nelle operazioni di analisi dei risultati di apprendimento** delle studentesse e degli studenti, non per sostituirsi all'insegnante, bensì per supportare le figure educative nel prevedere le probabilità che uno studente fallisca, interrompa o abbandoni la scuola (Ferro Allodola, 2021). L'AI, con la riflessione profonda del rapporto tra reale

e virtuale che porta nella scuola, può rappresentare **una nuova risorsa capace di promuovere e favorire l'inclusione di qualità**, grazie alle tecnologie multisensoriali, adottate per agevolare l'apprendimento di bambini, preadolescenti e adolescenti colpiti da disturbi dello spettro autistico, così come alle innumerevoli soluzioni per supportare studentesse e studenti con Bisogni Educativi Speciali.

Conclusioni

Sviluppare una cultura del digitale e ora anche una cultura dell'AI a scuola non può che trasformarsi nello sviluppo di un'apertura costante verso l'innovazione didattica, organizzativa, sociale. L'obiettivo è praticare una Digital Literacy allo scopo di promuovere una costante attività di sorveglianza della costruzione del pensiero critico (metacognizione e problem solving), **permettere una esposizione continua al confronto e alla messa in discussione dei pensieri propri ed altrui in relazione ad una frontiera di affidabilità**, sempre da discutere e confermare; imparare a confutare con scetticismo le teorie cospiratorie, continuando a credere nelle finzioni narrative ma perché contribuiscono a farci conoscere mondi possibili e non perché ci presentano soluzioni facili a problemi difficili (Gottshall, 2012); superare lo storytelling di una separazione tra scienza e humanitas per coniugarli invece sempre più strettamente e con creatività e fiducia.

C'è un tempo per esplorare in modo da prepararsi a comprendere, per non perdere l'occasione di migliorare. Il tempo è questo.

Bibliografia

Benanti, P. (2022). Human in the loop. Decisioni umane e intelligenze artificiali. Mondadori Università.

Calabrese, S. (2020). Neuronarrazioni. Editrice Bibliografica

Fabiano, A. (2022). Hypothesis for Better Social Justice: The Inclusive School between Digital Teaching and Artificial Intelligence. Formazione & Insegnamento, 20(1 Tome I), 116–126. https://doi.org/10.7346/-fei-XX-01-22_11

Ferro Allodola, V. (2021). L'apprendimento tra mondo reale e virtuale. Teorie e pratiche. Editzioni ETS.

Gottshall, J. (2014). L'istinto di narrare. Come le storie ci hanno reso umani. Bollati Boringhieri.

Kahneman, D. (2013). Pensieri lenti e veloci. Mondadori

Floridi, L. (2022). L'etica dell'Intelligenza artificiale. Sviluppi, opportunità, sfide. Raffaello Cortina Editore.

Ryan, M.L. (2004). Narrative across media: the languages of Storytelling. University of Nebraska Press.

Macchine in grado di fidarsi: le sfide del cognitive modeling

È possibile per un sistema di IA simulare attitudini umane come la fiducia facendo uso di architetture basate sul cognitive modeling? Proviamo a capire qual è il senso di questo approccio nell'era del machine learning

Di **Rino Falcone**, Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche, Roma

La sfida che **i sistemi di Artificial Intelligence** (AI) stanno rappresentando per la società risulta di enorme rilevanza: cresce la loro competenza e generalità di supporto e di conseguenza la pervasività in ciascun ambito delle nostre esistenze, individuale e sociale.

Si stanno diffondendo sistemi di AI che ai problemi loro sottoposti rispondono con risultati spesso del tutto indistinguibili da quelli che gli stessi umani realizzano. Al contempo, ci si confronta sui **dubbi** riguardanti le reali capacità intelligenti di questi modelli che seppure confortati da questi strabilianti risultati appaiono procedere secondo approcci quantitativi, tutt'altro che ovvi rispetto alle performance espresse.

Diventa perciò interessante chiedersi se un approccio guidato dal puro modellamento statistico e correlazionale dei dati sia sufficiente a riprodurre alcune attitudini tipiche degli umani nell'esercizio delle loro interazioni intelligenti. E quale senso abbia oggi costruire architetture, in modo top-down, che facciano uso del "**cognitive modeling**".

Presentiamo qui un approccio tipicamente basato sul modellamento cognitivo per la definizione e la conseguente simulazione di attitudini tipiche degli umani. In particolare, ci focalizzeremo su quella considerata base fondamentale della socialità: la fiducia. Proveremo a capire qual è il senso di questo approccio nell'era del machine learning.

Evoluzione dell'intelligenza artificiale

L'idea di fornire una macchina con capacità intelligenti analoghe a quelle degli umani, è un'idea che ha trovato un importante sviluppo verso la metà del secolo scorso, quando si introdusse una nuova area di ricerca, chiamata Artificial Intelligence. Anche se non possono essere dimenticati contributi precedenti fondamentali per questa area scientifica da parte di **Alan Turing** [1] e di **McCulloch e Pitts** [2], è nel 1956 che un fortissimo fermento di brillanti menti (John McCarthy, Marvin Minsky, Claude Shannon, Allen Newell e Herbert Simon e Nathaniel Rochester) avviano la sfida in modo ufficiale. Fu proprio McCarthy ad introdurre il nome di Artificial Intelligence [3].

AI forte e AI debole

L'approccio ha avuto differenti modelli e metodi ma si può sostenere che due principali direzioni di studio possono essere riconosciute: si tratta delle cosiddette AI forte e AI debole.

Nella AI debole la riproduzione del comportamento intelligente non risulta ispirata dal modo in cui gli umani lo realizzano: l'obiettivo è la **risoluzione del problema** e non l'analogia con il modo di risoluzione.

Nella AI forte, viceversa, l'obiettivo è di **realizzare la soluzione nello stesso modo in cui l'otterrebbe un umano**: il riferimento è quindi il confronto con i processi neuro-socio-cognitivi dei sistemi naturali. L'AI forte ha due ulteriori prospettive: l'analogia con i sistemi umani arriva all'idea di riprodurre le medesime strutture biologiche (e i conseguenti processi) degli esseri viventi: **prospettiva naturalista-connessionista**; oppure semplicemente **simulare la logica della psiche** (intentional stance [4]) degli esseri umani: prospettiva mentalista-cognitivista.

Questi differenti approcci hanno avuto vicende alterne, scambiandosi la prevalenza e il momentaneo (in)successo in fasi diverse. Comunque l'evoluzione delle tecnologie intelligenti è stata costante e i successi attuali sono incontrovertibili. Hanno giocato un ruolo fondamentale non solo l'elaborazione di tecniche strettamente di IA, come lo sviluppo di architetture cognitive sofisticate o gli algoritmi di apprendimento, ma anche una serie di evoluzioni scientifiche di campi più o meno affini alla AI: la potenza di calcolo, la miniaturizzazione dei componenti elettronici, gli sviluppi delle reti di comunicazione, la sensoristica sofisticata, lo sviluppo di nuovi materiali, la possibilità di costruire enormi basi di dati.

Tutto questo ha prodotto **una straordinaria trasformazione dei nostri ambienti commerciali, intellettuali e sociali**. Di fatto, determinando una profonda mescolanza dei mondi reali e virtuali in praticamente tutte le attività degli esseri umani. E di strumenti intelligenti di supporto in grado di fornire prestazioni crescentemente impattanti.

Tra gli straordinari avanzamenti di questi sistemi, ci sono quelli riconducibili in particolare alle tecniche di machine learning ed in particolare di deep learning. Si deve a queste tecniche il passo che conduce alla AI generativa (di cui parleremo più avanti).

Il modellamento cognitivo in IA

Per valutare il ruolo svolto dal c**ognitive modeling** per l'intelligenza artificiale, è necessario risalire alle ragioni del paradigma cognitivista evolutosi nel tempo e con natura fortemente multidisciplinare [5, 6] che fonda la comprensione del comportamento degli umani sulla capacità di modellare le loro rappresentazioni cognitive/mentali: rappresentazioni di cosa credono, intendono, vogliono, preferiscono. E come queste attitudini forniscano una interpretazione del mondo, degli altri, di essi stessi.

Al cuore di questo paradigma c'è quindi **la mente interpretata come un sistema in grado di operare su rappresentazioni**. È attraverso questa operatività della mente, ossia grazie alla sua capacità di manipolare, costruire, conservare, recuperare le rappresentazioni, che può essere spiegato il comportamento: per esempio come sarebbe possibile perseguire un fine/scopo senza averne prima una evidenza rappresentazionale? Lo stesso fine potrebbe certo essere raggiunto con altri mezzi, in modo inconsapevole o meglio senza rappresentarselo preventivamente, ma sarebbe legato a fattori casuali o a selezione evolutiva (come per molti animali a cui non è attribuibile una mente). Noi piuttosto siamo interessati a comprendere (anche allo scopo di simularlo) il pensiero

intelligente dell'uomo che evidentemente guida il proprio comportamento attraverso rappresentazioni finalistiche prestabilite.

Dato il ruolo pervasivo che svolgono oggi le tecnologie intelligenti, la loro capacità di interazione con gli umani in moltissime attività, sembra del tutto evidente ritenere che queste tecnologie siano dotate di attitudini particolarmente affini a quelle che hanno reso l'uomo l'animale sociale che conosciamo. Tra queste certamente c'è la fiducia: vediamone un'analisi concettuale e un modello, basato su cognitive modelling, che la rappresenta.

Il concetto di fiducia e il suo modello computazionale

Come vivremmo in un mondo in cui l'affidabilità degli altri fosse del tutto imprevedibile? In cui non avremmo modo di valutare se i compiti delegati agli altri, quelli che non siamo in grado di realizzare per nostro conto e da cui dipende il nostro benessere, la nostra salute o la nostra stessa sopravvivenza, fossero realizzati oppure no? Quali effettive relazioni si stabilirebbero tra gli individui? Che società si svilupperebbe? Ci sarebbe una società?

Questa premessa rende chiaro come la fiducia rappresenti un'attitudine essenziale nelle nostre vite e nei nostri comportamenti.

Anche per questo è stata, da sempre, oggetto di vasto approfondimento e studio in differenti ambiti scientifici: dalla filosofia, alla psicologia, alla sociologia, all'economia, alla biologia [7, 8, 9, 10]. Eppure, o forse proprio per questa vastità di punti di vista con cui può essere analizzata, non esiste una definizione condivisa e unica di fiducia.

Volendo cercare di delinearne i due caratteri principali, potremmo definirla:

come un ragionamento: essa implica infatti un modo di processare in modo razionale delle situazioni o delle caratteristiche o ancora dei dati, quindi di svolgere delle considerazioni logico-deduttive su questi elementi, individuare ipotesi considerabili oggettive e convincenti attraverso cui procedere ad un giudizio e quindi ad una decisione. Oppure,

come un sentimento: in questo caso entrando in gioco, non processi razionali ma piuttosto elementi di affettività ed emozionalità.

Fornendo una veste formale e operazionale al concetto di fiducia, possiamo dire che:

la fiducia **è uno stato e attitudine mentale**:

- ibrido: ossia tanto cognitivo, quanto affettivo;
- con struttura composita: riferibile a differenti ingredienti cognitivi: credenze, scopi, intenzioni, aspettative, etc.;
- orientato a differenti oggetti e dimensioni (ci si può fidare di un umano o di una sedia, per un compito ma non per un altro, e così via).
- **la fiducia è un fenomeno ricorsivo**: è possibile/necessario individuare delle ragioni per fidare e per ciascuna di queste ragioni è possibile/necessario individuare altre ragioni per fidarsi di esse stesse (e così via).

- **la fiducia è un processo sia mentale che pragmatico**. Mentale nel senso che può essere considerata tanto come una valutazione (una semplice attitudine mentale, una predisposizione, una valutazione preventiva non necessariamente connessa all'atto di fiducia); ma anche considerarla come una decisione (anche eventualmente dopo aver preso in considerazione comparazioni tra soggetti da fidare). Pragmatico, in quanto può essere considerata come un'azione (un comportamento, un atto intenzionale). In generale, è possibile pensare che uno sviluppo causale guidi i vari processi: per esempio che la valutazione, la decisione e l'azione siano rispondenti a stati mentali coerenti tra loro e, rispettivamente, ciascuno precondizione dei successivi (lo stato mentale della valutazione è predisponente lo stato mentale della decisione e lo stato mentale della decisione è predisponente lo stato mentale dell'azione). In realtà può succedere che questa coerenza non sia sempre rispettata.
- la fiducia è un fenomeno dinamico, non solo perché cambia nel tempo ma anche perché è possibile derivare fiducia da fiducia per esempio attraverso i fenomeni di transitività, o di categorizzazione, o dalla fiducia nelle credenze per fidare e così via.

Introduciamo **la seguente rappresentazione formale del concetto di fiducia**:

Trust(X Y τ C)

- d**ove X rappresenta il trustor**, l'agente che si affida, che sente fiducia; questo deve essere un agente cognitivo, ossia dotato di scopi e credenze interni ed espliciti;

- **Y è il trustee**, l'agente/entità che deve essere fidato; Y non è necessariamente un agente cognitivo, nel caso in cui lo fosse la relazione si connoterebbe come fiducia sociale.

- C è il contesto (l'ambiente) in cui il trustee deve operare per realizzare il compito delegato;

- **τ = (α,g) è il compito delegato**; esso corrisponde ad una coppia: azione (α) e stato del mondo (g); il legame della coppia è dato dal fatto che quell'azione è il mezzo per ottenere quello stato del mondo. Non sempre nella delega di un compito vengono esplicitate entrambe queste variabili. È possibile che il trustor deleghi direttamente lo stato del mondo g al trustee e il trustee poi decida come ottenere quello stato del mondo (ossia scegliere quale azione individuare per realizzarlo).

Data questa formulazione, Trust(X Y τ C), possiamo tradurla quindi sostenendo che l'agente (cognitivo) X che ha la necessità di ottenere un certo scopo, ossia una certa situazione nel mondo (lo stato g), deleghi all'agente (non necessariamente cognitivo) Y il compito τ. Ossia gli deleghi la realizzazione di una certa azione (α), nel contesto C, per fare in modo che si avveri lo stato del mondo g che è il suo scopo.

Quindi se ne deduce che affinché si abbia l'attitudine a fidare è indispensabile che il fidante (trustor) abbia degli scopi da perseguire. Di conseguenza, l'agente X deve essere un agente cognitivo. Questa constatazione porta con sé alcune conseguenze rilevanti. Per esempio, che per perseguire uno scopo è necessario fidarsi di qualcuno/qualcosa, al limite di sè stessi. Ma anche che il possesso (creduto dal trustor) di specifiche caratteristiche da parte di un trustee può attivare un potenziale scopo e la relativa relazione di fiducia (generare un nuovo scopo).

Si può inoltre pensare a forme generalizzate di fiducia, dove X può fidare Y per una certa tipologia di scopi (e solo per quella) o addirittura per "qualunque" scopo; o ancora, fidare un insieme di agenti per un dato scopo o per una famiglia di scopi.

Ci sono inoltre i cosiddetti **fenomeni di fiducia collettiva** che sono relati agli scopi (bisogni, aspirazioni) delle persone coinvolte.

Abbiamo quindi visto quale ruolo fondamentale svolge una delle componenti cognitive prese in considerazione: lo **scopo.** Di fatto non può esserci fiducia senza uno scopo da ottenere!

Ma ci sono altre componenti cognitive, che devono essere modellate e che hanno una grande rilevanza per comprendere e definire il processo di "fidare". Sono le credenze (beliefs) del trustor. Esse rappresentano le basi principali su cui la fiducia è fondata. Queste credenze sono rivolte principalmente al trustee ma non solo, per esempio riguardano anche il contesto in cui il trustee opererà. Vediamole nel dettaglio:

- una prima classe di credenze è rivolta alle competenze di Y; in particolare a quelle che sarebbero necessarie per il task o classe di task che X intende delegargli. Queste competenze generali possono essere ulteriormente specializzate: ci sono le abilità vere e proprie, ossia le capacità fisiche che un agente è in grado di esibire, ma anche il know-how, ossia le conoscenze utili per esercitare quelle abilità al meglio; e ancora la self-confidence, ossia la consapevolezza di quelle abilità (e così via).

- una seconda classe di credenze è rivolta alle intenzioni di Y; in particolare a quelle che sarebbero necessarie per il task o classe di task che X intende delegargli. Queste intenzioni si articolano in due differenti sottoclassi. Da una parte ci sono le attitudini intenzionali verso quel compito da parte di Y, indipendentemente da chi glielo ha delegato: la capacità di persistenza, motivazione, etc. direttamente collegabili a quel task o classe di task. D'altra parte, ci sono le attitudini intenzionali, attribuibili a Y sulla base della combinazione di quel task con la conoscenza di chi ha delegato il compito: benevolenza, non pericolosità, sicurezza, etc.

- una terza classe di credenze è rivolta alle conoscenze sul contesto in cui Y dovrà operare per realizzare il task delegato da X; ci sono infatti varie possibilità di condizioni favorevoli o di ostacolo alla realizzazione del compito. E l'azione di Y risulterà ovviamente influenzata da queste condizioni.

- un'ultima classe di credenze riguarda la dipendenza di X da Y per la realizzazione del task. In realtà questa dipendenza può essere di due tipi: X non è in grado di realizzare quel compito se non può delegarlo ad Y (dipendenza forte); oppure, per X è meglio delegare a Y piuttosto che svolgere da solo quel compito che pure sarebbe in grado di fare (dipendenza debole).

Ovviamente, quanto più è precisa la conoscenza di queste competenze, intenzionalità, contesti e dipendenze da parte di X, tanto più adeguata sarà l'aspettativa sui risultati delle azioni di Y.

Il modello illustrato assai sinteticamente nelle pagine precedenti, rappresenta **la base del modello socio-cognitivo della fiducia** [11, 12, 13].

A partire dallo schema mentale necessario per sviluppare l'attitudine di fiducia è possibile analizzare le molteplici caratteristiche ed effetti che derivano da questo concetto di base. Per esempio, quali sono le dinamiche della fiducia? Come si relaziona con l'ordine sociale ed in particolare con le norme e le autorità. Su quali tipologie di sorgenti si basa la fiducia? Come si misura la fiducia? Quale è il suo rapporto con l'autonomia e con il controllo? Come può essere sfruttata così da rappresentare un capitale utilizzabile in una rete sociale?

Rispondiamo, nel seguito, solo ad alcuni tra i più interessanti di questi interrogativi.

Le fonti di fiducia

Come abbiamo visto i comportamenti degli agenti, in particolare se essi si fidano o meno di altri agenti, dipendono da cosa essi credono su questi ultimi: ossia dalle loro relative beliefs. Queste credenze non sempre hanno lo stesso peso. Alcune possono essere più convinte, altre meno. Ma **da cosa deriva questa convinzione, ossia la forza di quelle credenze?** Essa deriva dalla fonte (o le fonti, se ce n'è più di una) che ha (hanno) permesso di generarle.

È quindi importante concentrarsi sulla **natura di queste fonti**, per comprendere appieno come le credenze si generino e si modifichino nel tempo. Una prima tipologia di fonte, la principale, è l'esperienza diretta: l'agente acquisisce la specifica credenza attraverso la percezione diretta (da parte dei propri sensi ma anche di alcune proprie capacità cognitive, tipo la memoria) del fenomeno che genera la credenza.

Una seconda tipologia di fonte è la **comunicazione** da parte di altri agenti, del fenomeno che genera la credenza. In questo caso si parla di esperienza indiretta, in quanto il fenomeno viene mediato da un altro agente che può aver esperito direttamente lui il fenomeno che genera la credenza o addirittura aver a sua volta ricevuto comunicazione indiretta del fenomeno.

Una terza tipologia di fonte è il ricorso da parte dell'agente primario (colui che si costruisce la credenza) a sue capacità cognitive particolarmente sofisticate come il ragionamento e l'inferenza. Attraverso queste capacità l'agente genera nuove credenze che possono derivare da esperienza diretta e indiretta ed essere elaborate. Esempi sono il ragionamento che permette la categorizzazione di elementi del mondo, piuttosto che il ragionamento per analogia.

Per ogni specifica credenza (belief), indipendentemente dalla natura della fonte che l'ha generata, è utile indicare alcune caratteristiche fondamentali del rapporto credenza-fonte:

- **identificazione della fonte**: è in grado l'agente che possiede la specifica credenza, di ricondurre quella credenza alla fonte che l'ha generata?
- **valore di certezza della fonte** sul contenuto trasmesso: la fonte che genera (o contribuisce a generare) la credenza ha fornito (o può essere dedotto) un valore di certezza sul contenuto?
- **fiducia verso la fonte**: considerata la fonte che potenzialmente può determinare la credenza di un certo agente, quale è la fiducia dell'agente verso quella fonte? E quanto è determinante per dare un valore alla credenza conseguente?

Come si vede c'è un interessante elemento di ricorsione nell'analizzare l'attitudine di fiducia. Per fidarsi infatti è necessario ricorrere a delle credenze, ma queste a loro volta hanno la necessità di essere considerate affidabili e quindi scatenare, da parte del trustor, un processo di fiducia all'indietro, verso le fonti (e poi le fonti delle fonti, e così via).

Dinamica della fiducia

Un esempio di dinamica della fiducia che dà conto della necessità di costruire un modello cognitivamente ricco e rispondente al profilo del trustee, viene dall'analisi del processo di fiducia, quando lo si confronta rispetto ad un modello semplificato.

Se supponiamo di avere il seguente schema (Figura2):

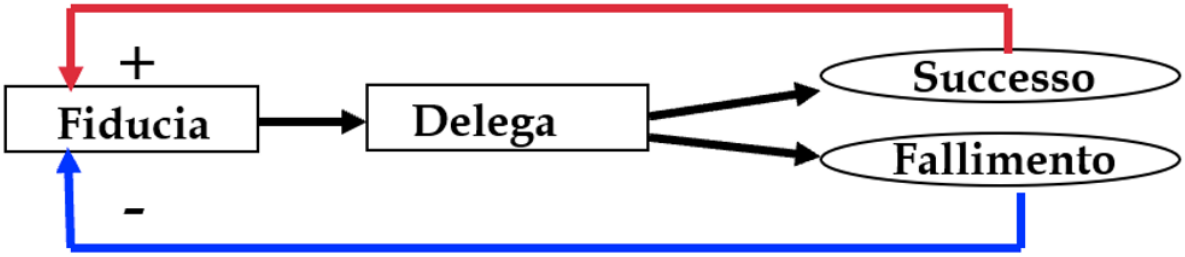


Figura2

Potremmo dire che ogni volta che grazie ad un livello di fiducia sufficiente si passa a **delegare un compito** e a valle di questo il trustee ottiene il risultato atteso (successo), la fiducia nel trustee si conferma o addirittura aumenta (linea rossa positiva di ritorno sulla fiducia). Insomma, il feedback è positivo e rafforza quel potenziale comportamento. Quando invece si ha un fallimento da parte del trustee, la fiducia del trustor in esso declina (linea azzurra negativa di ritorno sulla fiducia). Uno schema eccessivamente semplificato.

Se introduciamo un modello cognitivo del trustee, lo schema cambia (Figura 3).

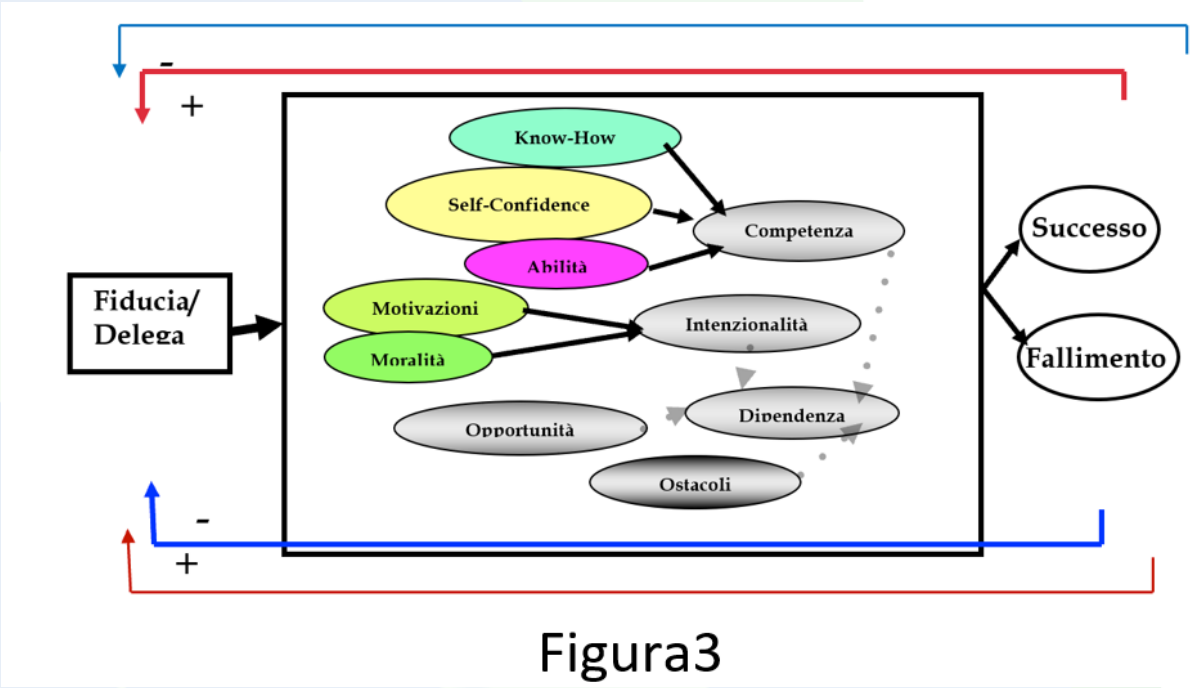


Figura3

In questo caso, come è riportato nella Figura3, sono associate tanto al successo quanto al fallimento del compito delegato, oltre ai feedback coerenti con il risultato (quelli già visti nella Figura2), anche feedback di segno opposto (in generale di peso inferiore). Questi stanno a rappresentare il fatto che nella valutazione della performance del trustee da parte del trustor è possibile, grazie alla valutazione dei vari fattori che entrano in gioco e che sono consapevolmente considerati ed eventualmente valutati dal trustor, anche elementi contribuenti in segno contrario al risultato ottenuto. Elementi che devono essere presi in considerazione per raffinare il modello del trustee e per adeguare la fiducia nei suoi confronti. Per fare un esempio, è possibile che la delega di un task al trustee veda una sua performance che realizza il task ma solo in quanto particolari condizioni

ambientali (non ordinarie) lo hanno favorito. Ed anzi nella performance si potrebbero rilevare deficit sulle sue competenze e/o sulle intenzionalità.

Insomma, una teoria attribuzionale, legata ad un modello cognitivo più sofisticato dell'attitudine a fidare, permette di andare oltre alla sola valutazione del risultato dell'atto di fiducia.

Un modello di fiducia analitico e articolato come quello sviluppato in [11, 12, 13] può quindi essere implementato in sistemi artificiali intelligenti (Figura4) così che l'interazione di questi sistemi sia più verosimile e adeguata a quella che gli umani esercitano nelle loro interazioni.

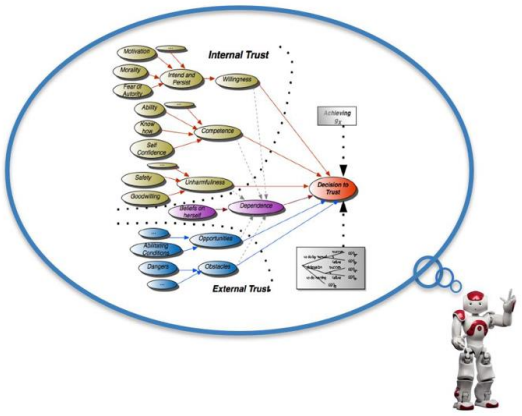


Figura4

Ma torniamo ora all'approccio non cognitivo e subsimbolico della AI.

L'AI generativa: potenzialità e dubbi

L'intelligenza artificiale generativa rappresenta quell'ambito di ricerca e studio che ha a che fare con la realizzazione di sistemi AI capaci di generare, a partire da dati precedentemente analizzati e appresi, nuovi dati o nuove versioni di dati esistenti. È noto il caso di ChatGPT [14], un modello generativo di testi capace di creare testi in risposta a domande. ChatGPT realizza questi testi partendo da una base informativa dello stesso genere. Esistono anche altri sistemi AI generativi che producono immagini, suoni o video, non sempre a partire da dati di analoga natura. La stessa OpenAI, l'azienda che produce ChatGPT , ha sviluppato DALL-E [15], un sistema che realizza immagini accedendo a descrizioni testuali esprimibili in linguaggio naturale oppure con un input misto testo e immagini.

L'AI generativa analizza enormi quantità di dati, e individua in questa massa di informazioni, schemi e regolarità (correlazioni statistiche) così da generare risultati statisticamente probabili. Può essere che il linguaggio espresso, è il caso di CHAT-GPT, risulti per lo più indistinguibile da quello prodotto da umani. Questo provoca sul giudizio degli umani che recepiscono quel linguaggio, un forte condizionamento nell'assimilare ad output simili, processi di produzione similari: in pratica a realizzare una similitudine tra pensiero umano e processi dell'AI generativa.

In realtà, al momento di pubblicazione di questo lavoro, sono stati individuati chiari limiti, incongruenze e fallimenti nelle performance di questi sistemi [16], discostando queste produzioni da quelle degli umani. Inoltre, **il modo in cui questi sistemi generativi funzionano è chiaramente differente dal modo in cui la mente umana ragiona e usa il linguaggio**. I primi sono

sostanzialmente motori statistici che macinando quantità impressionanti di testi, producono quindi testo con più alto valore di probabilità riferito agli schemi conversazionali appresi. La loro "comprensione" passa attraverso questa modalità di emergenza statistica (seppure molto sofisticata in questo ambito).

La mente umana, viceversa, si fonda su processi cognitivi ed affida la propria comprensione alla capacità di rappresentarsi il mondo e di operare su queste rappresentazioni, come detto all'inizio.

Sembrerebbe quindi che il modo di generare testo di questi sistemi AI, non rapportandosi esplicitamente al modello del mondo, non possa sviluppare del vero ragionamento, quello che permette la comprensione del mondo fisico e sociale basato su connessione di concetti ed entità. **In pratica, convergendo sul fatto che l'IA generativa permetta di apprendere implicitamente la sintassi (ossia le regolarità di forma con cui una lingua si esprime), sorgono dubbi profondi che sia in grado di apprendere la sua semantica (il significato delle cose che quella lingua esprime).**

Non è chiaro se e come questi limiti siano superabili e dipendenti non dal paradigma di base, quanto piuttosto dallo stato ancora non sufficientemente sviluppato della tecnologia che li realizza.

Va comunque sottolineato come l'indagine scientifica [17, 18], proceda nel tentativo di comprendere se l'efficacia di questi sistemi derivi esclusivamente dalla modellazione accurata delle statistiche di co-occorrenza di parole superficiali o se questi modelli rappresentino e ragionino anche sul mondo che descrivono. Questo genere di ricerca risulterà di grande importanza per comprendere meglio la natura profonda di questi algoritmi.

Conclusioni

Abbiamo quindi visto come l'approccio cognitivista e quello generativo all'IA offrano modelli e soluzioni interessanti ma non necessariamente convergenti al momento. Sembra inverosimile, sulla base dei limiti descritti, realizzare una AI generativa in grado di sviluppare una attitudine a fidarsi analoga a quella che può essere implementata in un sistema di AI con un'architettura cognitiva definita sulla base del modello presentato in questo lavoro. È possibile, d'altronde, che soluzioni miste di questi due approcci siano alquanto promettenti nella prospettiva di simulare i comportamenti intelligenti.

In questa chiave, visti gli avanzamenti e il livello impattante con cui queste tecnologie sono in grado di modificare i nostri più ordinari comportamenti -completamente trasformando il nostro modo di apprendere, di interagire, di operare nei vari ruoli che svolgiamo attivamente nel mondo- **è opportuno avviare una riflessione profonda sulla prospettiva che si presenta al genere umano dato lo sviluppo di questa tecnologia.**

Ci troviamo di fronte alla possibilità di straordinarie nuove opportunità ma al tempo stesso a potenziali e gravi rischi. Per esempio, la ormai diffusa digitalizzazione dei nostri comportamenti (quella che alimenta i big data, poi anche utilizzati dalla AI generativa) fa in modo che essi possano essere, nella maggior parte delle volte (a nostra insaputa o con la nostra benedicente sottovalutazione), osservati, memorizzati ed utilizzati. Questo solleva le ben note problematiche relative al cosiddetto "capitale di sorveglianza" [19]: il renderci produttori di dati che serviranno a realizzare previsioni particolarmente efficaci su futuri comportamenti, introduce i rischi a cui queste previsioni sono soggette (conformismo, riproposizione di schemi inadeguati seppur prevalenti. E così via). Inoltre, questo alimenta il potere delle industrie tecnologiche che hanno di fatto

monopolizzato il mercato mondiale sfruttando l'assenza di limiti e di regole, e che operano sostanzialmente senza assumersi responsabilità per gli eventuali usi distorti delle loro azioni e dei loro prodotti. Usi che possono produrre pericolose manipolazioni delle opinioni pubbliche e di alcuni principi essenziali delle libertà individuali, insomma un potenziale per lo stravolgimento delle governance delle società.

È per questo forte l'auspicio che le Istituzioni Pubbliche (nazionali e ancor più sovranazionali) preposte alla garanzia e tutela dei diritti fondamentali di ciascuno di noi, possano svolgere con determinazione ed equilibrio questo ruolo che compete loro. Affinché, controllando i rischi, non si debba rinunciare ai vantaggi di questa straordinaria sfida per l'evoluzione della nostra specie.

Bibliografia

[1] Turing, A.M. (2009). Computing Machinery and Intelligence. In: Epstein, R., Roberts, G., Beber, G. (eds) Parsing the Turing Test. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-6710-5_3

[2] McCulloch, W.S., Pitts, W. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5, 115–133 (1943). https://doi.org/10.1007/BF02478259

[3] John McCarthy & Claude Shannon (1958), Automata Studies, Journal of Symbolic Logic 23 (1):59-60.

[4] Dennett, D., (1987), The Intentional Stance, The MIT Press, Cambridge, Mass..

[5] Castelfranchi C., (2019), Fine della scienza cognitiva? Ma non del cognitivismo. Sistemi Intelligenti, 2019, 3, pp: 651-653 | DOI: 10.1422/95096.

[6] Nicoletti R., Rumiati R., Lotto L., (2017), Psicologia. Processi cognitivi, teoria e applicazioni, il Mulino, Bologna 2017;

[7] Luhmann, N. (1979) Trust and Power. Wiley, Chichester.

[8] Gambetta, D. (1988). Can We Trust Trust? In Trust: Making and Breaking Cooperative Relations. Blackwell. pp. 213-237.

[9] Jones, K. (1996). Trust as an Affective Attitude. Ethics, 107(1), 4–25.

[10] Hardin, R. (2002). Trust and Trustworthiness. Russell Sage Foundation.

[11] Falcone, R., Castelfranchi, C., (2001). Social Trust: A Cognitive Approach, in Trust and Deception in Virtual Societies, ed. by Castelfranchi C. and Yao-Hua Tan, Kluwer Academic Publishers, 55-90.

[12] Falcone, R., Castelfranchi, C., (2001), The Human in the Loop of a Delegated Agent: The Theory of Adjustable Social Autonomy, IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, Special Issue on "Socially Intelligent Agents - the Human in the Loop", 31, 406-418.

[13] Castelfranchi, C., Falcone, R., (2010). Trust Theory: A socio-cognitive and computational model. Wiley.

(14) https://openai.com/blog/chatgpt

(15) https://openai.com/product/dall-e-2

(16) Borji, A., (2023), A Categorical Archive of ChatGPT Failures. arXiv preprint arXiv:2302.03494v1, 6Feb2023.

(17) Li B., Nye M., Andreas J., (2021), Implicit Representations of Meaning in Neural Language Models, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 1813–1827.

(18) Li B., Chen W., Sharma P., Andreas J., (2023), LAMPP: Language Models as Probabilistic Priors for Perception and

(19) Zuboff, S., (2019), The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. London: Profile Books.

I vestiti nuovi dell'IA: storie di chat, test e tartarughe per andare oltre l'algoritmo

Proviamo a esaminare criticamente il nuovo immaginario IA e le aspettative degli utenti, proponendo una filosofia dell'IA centrata sulla rappresentazione computazionale dei saperi e del loro collante culturale e politico sommerso

Di **Ignazio Licata**, ISEM - Institute for Scientific Methodology, Palermo

L'**intelligenza artificiale** (IA) è forse la disciplina che ha più goduto degli **aspetti strategici dell'epistemologia**, alimentando un dibattito che si potrebbe riassumere in "ciò che L'IA fa" e "ciò che L'IA dichiara di poter fare" sulle possibilità di una convergenza asintotica tra naturale e artificiale. Oggi la questione si ripresenta con forza maggiore grazie all'Open IA (**ChatGPT** e **DALL- E2**) e al suo straordinario coinvolgimento collettivo.

Proviamo allora a **esaminare criticamente il nuovo immaginario IA** e le aspettative degli utenti, proponendo una filosofia dell'IA centrata sulla rappresentazione computazionale dei saperi del loro collante culturale e politico sommerso.

Le epistemologie hanno potere.

Hanno il potere non soltanto di trasformare i mondi

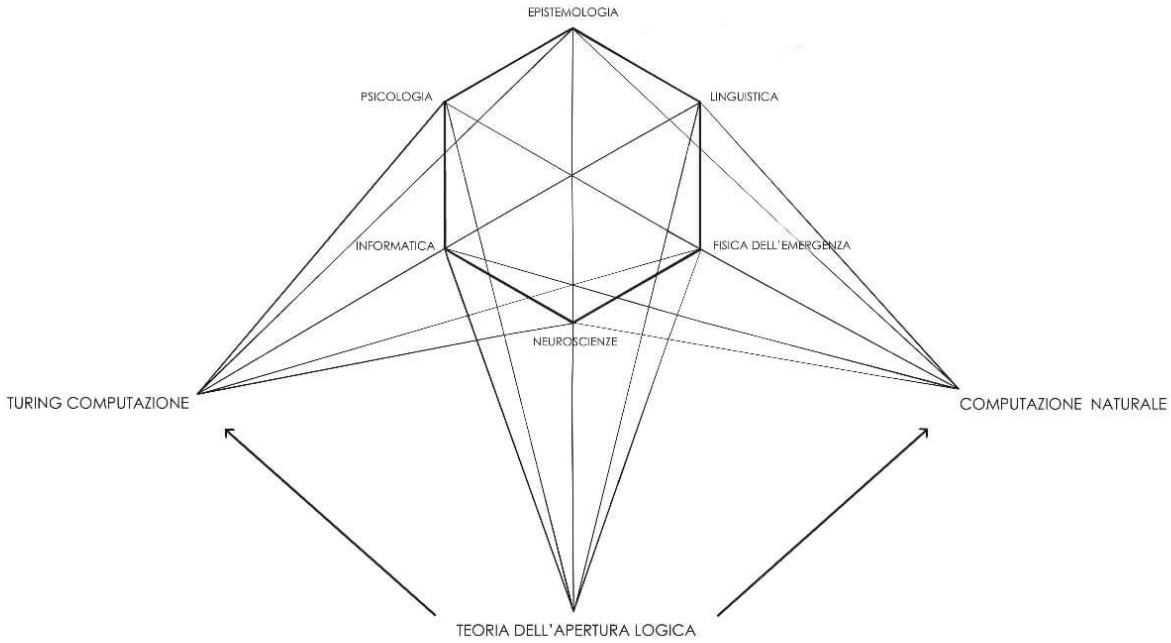
ma di crearli. (Nora Berenstain)

L'IA e i suoi confini sempre spostati in avanti

La visione tradizionale del lavoro epistemologico implica una riflessione critica sulla produzione dei saperi, sulla loro struttura e interpretazione, un'attività volta a chiarire la posizione di frammenti di teorie e pratiche all'interno di un tessuto stratificato, più o meno mobile, delle conoscenze, a fissare un rapporto tra la teoria e i fenomeni naturali e sociali. Si tratta in pratica di fornire quello che potremmo chiamare **un libretto di istruzioni ideale sugli aspetti cognitivi** e le possibilità di senso che la conoscenza ha in relazione al nostro stare al mondo. Un caso esemplare è quello della fisica quantistica, in cui il dibattito interpretativo è stato sin dall'inizio, ed ancora oggi, una forte esigenza interna della teoria che ha preso direzioni altamente formali e informa questioni fondazionali su "cosa c'è la fuori?" e sul "dicibile e indicibile".

Anche in discipline dove l'esito del dibattito epistemologico non conduce a risultati formali esiste **una dimensione interpretativa più o meno implicita**, o nascosta, che guida non soltanto l'uso di una teoria ma diremmo anche la sua reputazione. In **economia** la dimensione implicita è sempre fortissima, perché la scelta di variabili e parametri veicola (e vincola!) un'intera teoria delle relazioni umane e del rapporto uomo natura. La domanda che ci poniamo è se esiste qualcosa di simile per l'IA. Se fissiamo come data di nascita della disciplina il 1956, con i famosi seminari al Dartmouth College, bisogna ammettere che **l'IA nasce già con un forte bagaglio culturale fornito dal lavoro di A. Turing sul pensiero delle macchine e di J. Von Neumann** sulle analogie strutturali tra sistema nervoso e calcolatori (e forse si potrebbe risalire fin alla pascalina e all'homme machine): la filosofia dell' IA è, in generale, un'affermazione sulle capacità dei sistemi di elaborazione dell'informazione di emulare/simulare prestazioni umane.

Si tratta di affermazioni che implicano un modello della cognizione, al punto che **nel 1978 le pratiche dell'IA si fondono culturalmente con la più vasta area delle scienze cognitive**. Nel 1956 invece gli assunti sulla cognizione umana (M. Minsky: il cervello è una macchina di carne) venivano celati da dichiarazioni programmatiche di iper-performatività computazionale, ritenuti più adatti a catalizzare consensi e finanziamenti. La vecchia IA forte ha sostenuto per un buon numero di anni ruggenti un isomorfismo tra mente e macchina- di volta in volta posizionato a livello simbolico o subsimbolico-, che a ben vedere oggi appare piuttosto ingenuo, anche senza arrivare alla critica della ragione artificiale di Dreyfus o al dibattito tra R.Searle e i coniugi Churchland (Dreyfus, 1992; Searle vs Churchland, 1992) Si arriva così ad una IA debole, connessa alle altre discipline nel "diamante" delle scienze cognitive (fig. 1), e criticamente più consapevole dei limiti dei propri modelli.



Il diamante delle scienze cognitive, da Ignazio Licata, La Logica Aperta della Mente, Codice, 2008

Va detto però che in tutta la sua storia l'IA ha mostrato **una forte tendenza alla mitopoiesi e allo slittamento nel futuro dei propri obiettivi/promesse** che ne costituisce la filosofia nascosta ma pervasiva, ampiamente indipendente dai successi tecnici al punto da rivelarsi una sorta di propaganda. In fondo è questo che ormai ci si aspetta dall'IA: un confine che si sposta sempre in avanti, affascinante come l'Uomo Bicentenario (C. Columbus, 1999) e inquietante come il Proteus del romanzo di Dean Koontz (1973), trasposto in film qualche anno dopo da D. Cammell, Generazione Proteus (1976). La singolarità di R. Kurzweil è forse l'immagine archetipale di questa soglia critica (non definita) in un futuro (indeterminato) che vedrà l'IA diventare un'entità autonoma e imprevedibile.

Cosa accade in questi giorni in cui tutti parlano di friendly AI (ChatGPT, Dall-E 2)?

Si potrebbe rispondere indifferentemente: molte cose oppure veramente molto poche. Dipende da quale prospettiva osserviamo l'interazione tra gli utenti con i nuovi prodotti IA.

Il Test di Turing e le tartarughe di Valentino

Alan Turing possedeva un genio speciale nel tagliare, o almeno semplificare, i nodi gordiani entro i quali altri studiosi restavano intrappolati. La Macchina di Turing, esemplare fusione di astrazione formale e concretezza meccanica, traduceva le difficili questioni logiche di K. Godel e A. Curch in termini computazionali definendo un nuovo stile di lavoro sulle questioni fondazionali della matematica. La sottile questione della decidibilità diventava: la macchina si ferma o no? Non meno brillante è **il famoso test**, che mirava a porre la difficile definibilità dell'intelligenza in termini che i fisici direbbero operativi, i.e. data un'interazione dialogica a distanza, ad esempio tramite tastiere, tra un essere umano A e un sistema artificiale X, quest'ultimo sarebbe stato definito intelligente se tale sarebbe apparsa la sua performance all'interlocutore A. In questa semplice situazione sono molte le finezze, tipiche dello stile di Turing, su cui vale la pena soffermarsi.

Innanzitutto, va osservato che **il test non fa altro che riprodurre ciò che effettivamente accade quando due esseri umani si incontrano**. Entrano in gioco, dunque, il tempo di interazione, gli argomenti del dialogo e la soggettività del giudizio. Quest'ultima può essere tradotta in una probabilità secondo l'accezione della scommessa di Bruno De Finetti, ossia **una valutazione numerica di verosimiglianza assegnata dal valutatore** (De Finetti, 2006). E' chiaro che con l'aumento del tempo di interazione e degli argomenti la puntata della scommessa potrà subire variazioni positive e negative. In questo modo si evita una generica definizione di intelligenza e si sposta il giudizio su un piano relazionale.

La questione si può riassumere nella domanda: **rispetto a chi X è intelligente?** Nei primi '80, quando ero uno studente di fisica, l'avvento dei personal computer e la febbre della programmazione ci portò completamente fuori rotta rispetto agli studi. Ricordo alcuni mesi in cui il mio unico intento era quello di **realizzare un programma per produrre haiku** (con risultati poche volte ottimi, quando il caso agisce in modo illuminante, ma solitamente deludenti) e programmi di conversazione ispirati alle performance ormai famose di **ELIZA**. Si trattava di un modello di psichiatra rogeriano messo a punto tra il 1964 e il 1966 da Joseph Weizenbaum, al quale nel 1972 lo psichiatra Kenneth Colby oppose PARRY, un modello di paranoico.

I dialoghi tra i due, come si può immaginare, hanno il sapore di **un frammento infelice di Jonesco**. Questi esempi dovrebbero farci comprendere che **quello che si discute oggi in relazione all'Open IA non è nato in una notte**, per citare una frase di Jeff Jonas a proposito dei Big Data. Quello che oggi fa la differenza è l'uso brillante di molte infrastrutture software create per i social. I modelli di Open IA non sfuggono alla legge universale dell'informatica (GIGO: garbage in , garbage out) ma possono aggirarla in un gran numero di modi grazie ad una quantità di risorse informazionali pari, idealmente, all'intera rete.

E' questo che produce l'effetto straniante che molti utenti ricevono dai nuovi oracoli in sessioni che potremmo chiamare un gioco di Turing globalizzato. Eppure **non mancano lavori che mettono bene in evidenza i sempiterni limiti** di programmi che adesso scrivono buoni testi in autonomia e riescono a gestire conversazioni assai più articolate e sensate rispetto ai loro nonni ELIZA e PARRY (Floridi & Chiriatti, 2020). Come spiegare allora l'entusiasmo generale?

La storia delle tartarughe robot

La storia delle tartarughe robot può aiutarci a trovare una risposta. Nel suo libro I veicoli pensanti. Saggio di psicologia sintetica (prima ed.1984,2008) **Valentino Braitenberg, uno dei maggiori esponenti della cibernetica in Italia**, descrive i suoi esperimenti con dei piccoli robot, somiglianti

a tartarughe, in cui aveva incrociato sensori e attuatori in una sorta di chiasma artificiale, termine che in neuroanatomia si riferisce ad uno schema di fasci nervosi incrociati largamente presente nelle forme viventi. I sensori erano delle fotocellule, ma la dinamica che ne derivava era definita dal modo in cui il chiasma era realizzato. **Una tartaruga si avvicinava velocemente alla luce per sbatterci contro o si fermava in prossimità**, un'altra evitava la sorgente luminosa, oppure si avvicinava per poi allontanarsi rapidamente via. Per un osservatore esterno, ignaro della circuiteria, questi comportamenti potevano essere descritti facendo riferimento ad attitudini cognitive tipicamente umane: amore, odio, curiosità, paura.

È evidente l'analogia con il test di Turing e la scommessa di De Finetti: da una parte c'è **un comportamento meccanico** (come si diceva una volta: un mero comportamento meccanico), **dall'altra un osservatore che esprime un giudizio soggettivo** su questi comportamenti. Ancora una volta, è una situazione che può applicarsi alle relazioni umane, nessuno di noi conosce il collegamento tra comportamenti e correlati neurali (nostri o altrui), e il giudizio, più che riflettere una realtà "oggettiva", è il risultato di una "scommessa" cognitiva. Quello che accade con l'Open IA è l'incontro tra una nuova generazione di prodotti software e gli utenti social, già da tempo normalizzati per interagire con programmi a cui attribuiscono non soltanto intelligenza, ma addirittura consapevolezza e capacità artistiche. E come nel test di Turing originale queste performance ci dicono più sugli utenti che sulle strabilianti caratteristiche di questa nuova generazione di IA.

Come siamo arrivati a questo punto

In ogni articolo o saggio sull'IA arriva il momento di un po' di pedanteria che fatalmente tende a smorzare gli entusiasmi, ma che permette di vedere più a fondo dentro "le tartarughe". **I generatori di testi dell'Open IA sono modelli linguistici autoregressivi** basati su reti neurali con un altissimo numero di parametri che lavora su una memoria molto grande. Non ci interessa in questa sede citare i numeri oggi in gioco, è verosimile che diventeranno molto più grandi tra non molto. Nella fase di addestramento la rete regola i suoi pesi in modo da poter **stabilire la massima probabilità di connessione**, ad esempio, tra una frase e la parola che segue. Si tratta dunque di uno strumento statistico che può agire a vari livelli sul testo richiesto dal prompt dell'utente. Senza troppo sforzo si può definire un'architettura simile "biomorfa", almeno rimanendo in uno schema hebbiano d'antan come poteva essere ai tempi di Von Neumann. Quanta intelligenza può raggiungere un sistema di questo tipo?

Arriva adesso la parte pedante. L'ostacolo ai sogni di quella che si chiamava IA forte era dato dalla bassa apertura logica dei sistemi (Licata, 2018). In altre parole, **i programmi IA funzionavano molto bene in ambiti semanticamente chiusi, come gli scacchi, dove il movimento del pezzo coincide con il suo significato**. Quando si andava in contesti con maggiore apertura logica, dotati di una gamma di significati meno univoca, cominciavano ad intravedersi quel tipo di falle che ognuno di noi ha incontrato nella traduzione automatica. Abbiamo poi preso atto che **le traduzioni in rete miglioravano sensibilmente**, e se questo avveniva era perché erano già entrati in gioco alcuni di quei sistemi di parsing che oggi sono confluiti nell'open IA.

Quello che le nuove forme di ingegneria del software sono riuscite a fare è di riuscire a trattare la richiesta e l'elaborazione del testo richiesto dal prompt come fosse un sistema semanticamente chiuso, ossia di risolvere una quantità di testo sino a pochi anni fa impensabile in termini sintattici utilizzando forme piuttosto raffinate di gerarchizzazione dei testi reperiti nelle fonti.

La semantica può essere ridotta a sintassi?

Il risultato è notevole, e sposta la questione su un altro piano: la semantica può essere ridotta a sintassi? Ricordiamo che la cognizione umana lavora in un certo senso au contraire, si parte da una richiesta di senso applicata ad una selezione di dati ed eventi, si ipotizzano connessioni e soltanto alla fine si arriva ad una produzione linguistica e testuale, procedimento con un ampio margine di arbitrarietà soggettiva e incertezza che può condurre in fallo (esemplari sono le disavventure epistemiche dell'alter ego dello scrittore in Cosmo di Witold Gombrowicz (Gombrowicz,2004), ma che in altri casi realizza quello che impariamo a individuare come stile e poco ha a che fare con il tessuto esplicito della narrazione, la trama. Naturalmente, **dato un autore, lo stile può essere esaminato analiticamente** (in fondo è anche questo che fanno critici e filologi), e si scoprirà che **è dato da un certo numero di fattori**, come l'uso ricorrente di situazioni e immagini, il modo di mettere in rilevo il peso di un termine all'interno di un brano e così via.

Quello che può essere analizzato può essere modellizzato e simulato, e non è escluso che **in futuro le nuove forme di IA potranno produrre testi articolati con un sapore dato**. Arrivati a quel punto la sfida già in corso tra programmi generatori di testo e programmi che individuano l'artificiale, in una nuova versione del test di Turing, diventerà assai ardua, e va detto che per il momento il punteggio è decisamente a sfavore degli smascheratori. Ma la questione sintassi/semantica difficilmente può essere risolta da future performance. Il motivo è che, da un punto di vista sistemico, la cognizione umana è un amplificatore di informazione in virtù di processi ininterrotti di rottura di simmetria che corrispondono all'**emergenza di nuovi domini di significato**; questo implica forte dissipazione (una distruzione di gran parte dell'informazione pregressa accumulata) e avviene in interazione con l'ambiente.

In questo quadro **la memoria è una funzione dinamica, non un deposito passivo di informazione conservata**, e questo è il motivo per cui il miglior modello della cognizione umana è offerto dalla super-rete neurale del Quantum Brain in cui i nodi si distruggono e si accendono continuamente, modello che utilizza un formalismo preso a prestito dalla teoria quantistica dei campi (Vitiello, 2001; 2008). E' interessante ricordare che un modello di questo tipo permette una descrizione dei qualia (Humphrey, 2007), stati soggettivi legati al rapporto dinamico tra corpo e ambiente, e questo ci rimanda alla grande lezione dell'embodied cognition per cui non soltanto non c'è cognizione senza la complessità di un corpo immerso in un ambiente(Gibson, 2014), ma la dimensione emotiva- al di là delle chiacchiere new age- non può più essere considerata una nebbia che offusca la ragione cristallina, ma di quest'ultima è il motore da cui emerge l'intenzionalità. Durante una conferenza pubblica l'autore si è visto opporre come esempio della consapevolezza dell'IA un testo sulla paura. L'algoritmo aveva svolto onestamente il suo lavoro e aveva concluso che per una macchina la paura consisteva nell'essere spenta! L'interlocutore del pubblico, forse inconsapevolmente, aveva evocato una mezza dozzina di questioni filosofiche su riferimento e denotazione, ma una descrizione della paura non è avere paura.

Tanto basta per la nostra dose di pedanteria, il cui obiettivo non è quello di stabilire oggi cosa potrà fare domani un possibile algoritmo definitivo (Domingos, 2016), ma suggerire di tenere a mente che **un elaboratore di testi non è un sistema ad alta apertura logica**, anche se la reazione dell'osservatore può andare in altro senso (Sipper et al. 1999). Piuttosto, in epoca di forte omologazione verso il basso, va tenuto presente il monito espresso da Walter Siti in un suo recente articolo dal titolo: La società in cui gli scrittori pensano come ChatGpt (Domani, 27 Febbraio 2023). Ci siamo concentrati sui testi e non abbiamo parlato di **Dall-E2** perché in quest'ultimo scenario i vestiti nuovi mostrano la loro totale assenza e il giochino è fin troppo scoperto. Assieme ad una generale lontananza dalla comprensione del fatto artistico da parte degli utenti, per cui,

parafrasando Siti, si potrebbe dire: la società in cui i non artisti pensano di poterlo diventare con Dall-E2.

La sfida e le benefiche virtù dell'incertezza

Torniamo all'inizio del nostro percorso, alle questioni epistemologiche, non prima di aver dissipato l'impressione evocata verosimilmente in qualche lettore di essere autori di articolo "critico" nei confronti delle sorti magnifiche e progressive aperte dalle nuove forme di IA. Non è così.

Ci siamo concentrati sul machine learning per il natural language processing perché è quello che sta ottenendo la più alta attenzione mediatica, cosa del resto comprensibile poiché la comunicazione e lo scrivere nel web sono ingredienti costitutivi della nostra natura quanto la scrittura tradizionale. Inoltre ci ha offerto un'idea semplice della caratteristica davvero rivoluzionaria di questo modo di fare IA: per la prima volta, in modo evidente per tutti, si è compreso che **l'IA non è un qualcosa che sta dentro una scatola magica** (software o hardware), **ma una risorsa di rete estremamente complessa**, che emerge dall'interconnessione a più livelli di una molteplicità di agenti, uno scenario non troppo dissimile da quello delineato da Marvin Minsky anni fa in **La società della mente** (Minsky, 1986), e da F. Heylighen alla fine degli anni '90 (Heylighen e Bollen, 1996).

Non si tratta di un'intelligenza simile a quella umana (molti umani hanno sempre più difficoltà a superare il test di Turing nel riconoscere l'algoritmo e l'artificiale), e le questioni della embodied cognition, dell'autorappresentazione e la coscienza seguono per il momento altre linee teoriche (Chella ed al. 2008; Chella e Manzotti, 2007), ma al di là del clamore, i risultati sono di grande rilevanza e **oggi si può soltanto intravedere il potenziale impatto sulla vita e il lavoro.**

Pensiamo a tutto quello che può essere standardizzato attraverso le procedure informatiche, dalla selezione del personale all'accesso al welfare, l'allocazione di risorse, le procedure giuridiche e l'attivazione di sistemi d'arma(ma tante altre se ne potrebbero citare). **L'idea oggi dibattuta della giustizia predittiva è contenuta già in nuce nella prima cibernetica** (Wiener, 1968) ed è basata su premesse "molto ragionevoli": il recupero di leggi e sentenze, la messa a punto di strategie legali (il caso recente di DoNotPay), la produzione di atti e l'archiviazione, ma appena al di là dell'efficienza si trovano questioni spinose, emblematiche dell'intera sfera dall'IA applicata ad aree sensibili della vita umana. **La questione della recidività di un condannato**, ad esempio, difficilmente può essere delegata ad un sistema artificiale, rischiando così di diventare una caricatura di **Minority report** (Spielberg ,2002 da PK Dick, 1956). Infatti esiste per giustificare questa cautela una ragione di carattere generale, il teorema di Arrow sull'impossibilità delle scelte univoche dove ci sono interessi plurali e contrapposti, che può essere ulteriormente declinato tenendo conto che sistemi fortemente interconnessi possono mettere assieme una pluralità di ritratti di ciascuno di noi, a seconda dei tipi di interazione che abbiamo in rete, ma difficilmente ci riconosceremo in quelle rappresentazioni "zippate" e parziali; tra l'una e l'altra ci sono più cose in cielo e in terra di quante non ne conosca la filosofia dell'algoritmo.

Non si tratta dell'ennesima riproposizione del mito dell'ineffabile umano (almeno per quest'autore), ma di riconoscere che i **processi cognitivi non sono meri sistemi dinamici**, hanno una storia fatta di biforcazioni, sovrapposizioni, crisi, mutazioni, interpretazioni e scommesse individuali ed è su questa linea d'ombra di complessità che ogni giudizio, a cominciare da quello del diritto, tende a conservare un principio di ricerca del giusto nel legale (Romano, 2012; 2018; Miceli, 2023). Si potrebbe obiettare che in tal modo si abdica ad una giustizia trasparente per restare all'interno dell'errore e dell'incertezza, obiezione che potrebbe essere accolta come corretta, ma va controbilanciata all'interno di **una consapevolezza più ampia che riguarda ogni forma di**

giudizio umano, quella di essere figlio del tempo, sovradeterminato dai paradigmi valoriali, culturali e politici, e sotto determinato dalle vicende individuali contingenti che convergono nella questione da giudicare, radicato in conflitti permanenti e indecidibili (Star e Bowker, 2007).

L'incertezza diventa in questo caso la possibilità di porsi al di fuori e prima di ogni "certezza" in modo da reinstaurarla o ricostruirla da zero. Questo modo di pensare può trovare conferma proprio nell'universo della matematica, all'interno della quale provengono gli algoritmi. Appare infatti sempre più chiaro, a partire dal dibattito sui fondamenti dopo la lettura di G. Chaitin dei teoremi di Gödel, che **la matematica è un sistema aperto, soggetto a processi tellurici di riassestamento**, possibilità interpretative, e biforcazioni di sviluppi formali. Questo perché i matematici non sono manipolatori di simboli, ma assegnano a quei simboli un senso ed è questo che rende l'attività del matematico affascinante, difficile e bella (Lolli, 2022; Longo, 2021).

Pur restando lontani da giustificazioni fondazionali, va notato che **i nuovi modelli di IA di rete, a causa della loro complessità, rendono estremamente opache le procedure di selezione, categorizzazione e gerarchizzazione dei dati**, le scelte di ragionamento e dunque la formazione dei concetti che stanno alla base delle performance del sistema.

Attualmente nell'Open IA confluiscono dati istituzionali e privati, in parte anche ottenuti da procedure di marketing, ed è verosimile che per molti aspetti le cose resteranno così, compatibilmente con le direttive economiche che regolano il rapporto tra istituzioni e aziende. **Questa opacità nasconde il vero status epistemologico dell'IA**, quello di un**'ingegneria globale della conoscenza** che sistematizza fatti, teorie e scelte politiche (priorità culturali, direttive economiche, definizione di classi di soggettività, tutele sociali), nella gigantesca opera di costruire una rappresentazione del mondo delegata all'autorità di un oracolo infallibile e con l'impatto sociale di un golem. Per contro, è qui che **un'epistemologia dell'IA si definisce come una critica permanente della rappresentazione algoritmica dei saperi**, un'analisi delle ideologie implicite, attività che si svolge nel territorio dell'incertezza e del conflitto, prima di ogni scelta e categorizzazione (Mitchell, 2022; Numerico, 2021; Crawford, 2021).

Conclusioni

Nel 2017, in una delle sue ultime apparizioni pubbliche, Stephen Hawking tornò su uno dei suoi temi favoriti, la pericolosità di un effetto Proteo da parte di un'IA sempre più raffinata e articolata. Se nella sua forma letterale la profezia di Hawking appare quanto meno assai lontana, va considerata la versione più pragmatica che ne ha offerto recentemente H. Kissinger (Kissinger et al.,2022). Già oggi, assai prima di ogni singolarità, è ipotizzabile che **gruppi di cyberterroristi possano gestire risorse IA in modo destabilizzante**, questione che si aggiunge al già gravoso carico di un mondo complesso (Floridi, 2022; Tehrani, 2021). Un certo gusto trascendente e apocalittico ha sempre fatto parte della narrazione dell'IA centrata sui futuri possibili, ma la presenza ormai pervasiva delle nuove forme di IA, la loro straordinaria potenza e la crescita accelerata ci interrogano culturalmente e politicamente su questioni che si trovano al di là dell'algoritmo.

Bibliografia

Braitenberg, V. I veicoli pensanti. Saggio di psicologia sintetica, Mimesis, Milano- Udine, 2008

Chella, A. Integrazione, autoadattamento e coscienza artificiale, Sistemi Intelligenti, 3, 2008

Chella, A., Manzotti, R. (Eds) Artificial Consciousness, Imprint Academic, 2007

Crawford, K. Né intelligente né artificiale. Il lato oscuro dell'IA, Il mulino, Bologna,2021

De Finetti, B. L'invenzione della verità, Raffaello Cortina, Milano, 2006

Domingos, P. L'algoritmo definitivo, Bollati Boringhieri, Torino, 2016

Dreyfus, H. What Computers Still Can't Do: A Critique of Artificial Reason, MIT Press, 1992

Floridi, L., Chiriatti, M. GPT-3: Its Nature, Scope, Limits, and Consequences, Mind and Machines, 30, 681–694 (2020)

Floridi, L., Etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide, Raffaello Cortina, Milano, 2022

Gibson, J. J. L'approccio ecologico alla percezione visiva, Mimesis, Milanoi- Udine, 2014

Gombrowicz, W. Cosmo, Feltrinelli, Milano, 2004

Heylighen, F. , Bollen, J. The World-Wide Web as a Super-Brain: from metaphor to model, in Cybernetics and Systems '96 R. Trappl (ed.), Austrian Society for Cybernetics, 917-922, 1996

Humphrey, N. Rosso. Uno studio sulla coscienza, Codice, Torino, 2007

Kissinger, H., Schmidt, E., Huttenlocher, D. The Age of AI: And Our Human Future, Hodder And Stoughton, 2022

Licata, I. La logica aperta della mente, Codice edizioni, Torino, 2018

Lolli, G. Matematica in movimento. Come cambiano le dimostrazioni, Bollati Boringhieri, Torino, 2022

Longo, G. Matematica e senso, Mimesis, Milano-Udine, 2021

Miceli, M. Il processo artificiale, Un ragionevole dubbio sugli algoritmi in tribunale, Divergenze, Pavia, 2023

Minsky,M. La società della mente, Adelphi, Milano, 1986

Mitchell, M. L'intelligenza artificiale. Una guida per esseri umani pensanti, Einaudi, Torino 2022

Numerico, T. Big data e algoritmi. Prospettive critiche, Carocci, Roma 2021

Romano, B. Algoritmi al potere. Calcolo giudizio pensiero, Giappichelli, Torino 2018

Romano, B. Forma del senso. Legalità e giustizia, Giappichelli, Torino 2012

Searle vs Churchland: Searle, J. R., La mente è un programma?; Churchland, P.M. & P.S., Può una macchina pensare? In Mente e macchina (a cura di G. Lolli), Quaderni Le Scienze, 1992

Sipper,M. Ronald, E. M. Capcarrère, M. S. Design, observation, surprise! A test of emergence, Artificial Life, 5(3):225-3, 1999

Star, S. L., Bowker, G. C. Enacting silence: Residual categories as a challenge for ethics, information systems, and communication, Ethics and information Technology, 9, 273–280 ,2007

Tehrani, P. M., Cyberterrorism: The Legal and Enforcement Issues, World Scientific, 2021

Vitiello, G. Essere nel mondo: io e il mio doppio, Atque, 5, 155-176 2008

Vitiello, G. My Double Unveiled: The dissipative quantum model of brain, John Benjamins Publishing, 2001

Wiener, N. La cibernetica: Controllo e Comunicazione nell'animale e nella macchina", Il Saggiatore, Milano, 1968

Generative AI, dov'è il bene per l'Umanità?

Dobbiamo per forza continuare su una traiettoria di ricerca verso sistemi di IA sempre più potenti? Quali strumenti attuare per ovviare all'attuale quasi-monopolio cognitivo delle big tech? Come evitare una collettiva hallucination e preservare il nostro senso critico? Rischi e conseguenze della nuova strada imboccata dall'IA Generativa

Di **Mauro Lombardi**, Scienze per l'Economia e l'Impresa, Università di Firenze

La **Generative Artificial Intelligence** sta imboccando una nuova strada. Con la diffusione degli LLMs, il grande successo di **ChatGPT**-3 e il lancio di GPT-4 stiamo forse entrando in un'era contraddistinta da ciò che Floridi e Chiriatti (2020) chiamano "industrial automation of text production". Un'**automazione industriale della produzione di testi che trasforma radicalmente la scrittura umana**, supportata da strumenti capaci di combinare frammenti tratti da basi informative eterogenee in rappresentazioni linguistiche simili a quelle umane.

Soprattutto per coloro che svolgono professioni basate sulla scrittura (ma in realtà vale per tutti), il "cut & paste" potrebbe essere progressivamente sostituito dal "prompt & collate" (Floridi e Chiriatti, 2020: 691). Ecco con quali rischi e conseguenze.

https://www.agendadigitale.eu/cultura-digitale/un-mondo-senza-piu-creativita-per-colpa-dellai-una-proposta-per-scongiurare-il-rischio/

L'avvento dei LLMs (Large Language Models) e di ChatGPT

Nell'odierno scenario tecno-economico, è difficile distinguere tra realtà effettuale e rappresentazione immaginifica. Il 2023 è iniziato all'insegna del GPT-3 e delle sue notevoli performance, ottenute sulla base di alcuni elementi costitutivi. il funzionamento si avvale di **175 miliardi di parametri e 45 terabyte di dati desunti da testi**, il cui ammontare è stato stimato pari a **un quarto della Biblioteca del Congresso USA e a circa 300.000 metri lineari di libri**.

Sono stati spesi **12 milioni di dollari** per un lungo processo di addestramento del modello sull'enorme set di dati. Ha la capacità di creare output linguistici da combinazioni di testi e immagini per una varietà di compiti, richiesti dagli utenti, senza che gli sia sottoposto un esempio in precedenza (in gergo "one-shot fashion").

GPT-3 è il più potente modello di linguaggio mai costruito e ha mostrato di saper fare cose strabilianti:

- scrivere codici fino a generare brividi a John Carmack, pioniere della 3D computer graphics, citato da Heaven (2020);
- elaborare lo scritto "L'importanza di essere su Twitter", innescato da Mario Klingerman, artista che lavora con ML e si è trovato di fronte ad un output nello stile dello scrittore Jerome K. Jerome;
- c'è anche un articolo su GPT-3 scritto dallo stesso GPT-3: "OpenAI's GPT-3 may be the biggest thing since bitcoin", 18-7-2020.

A questi episodi suggestivi si potrebbe aggiungere il fatto che, alla domanda "se GPT sia l'App con il più alto tasso di crescita della storia". la risposta è stata: GPT non è un app, bensì un sistema di Machine Learning. Queste e altre prestazioni non convincono però Heaven (2022) che, pur ritenendo che GPT sia "shockingly good" enfatizza come esso sia ben lontano da una vera intelligenza, né Gary Marcus (2020), il quale chiarisce che GPT "non ha la minima idea di cosa stia parlando". Il successo, conseguente alla scelta di rendere accessibile al pubblico ChatGPT dal Novembre 2022, è stato rilevante: **100 milioni di utenti in gennaio, solo due mesi dopo il suo lancio, mentre per raggiungere lo stesso numero sono occorsi 9 mesi a TikTok, cinque anni a Google e Facebook** (Tung, 2023).

I risultati della Generative Artificial Intelligence

Una ragione fondamentale della rapida conquista dell'immaginario di così tante persone, da parte di ChatGPT, è che siamo di fronte a **Generative Artificial Intelligence (Generative AI)**. Ottiene risultati che appaiono creativi, perché le combinazioni di testi e immagini sono realizzate mediante l'introduzione di **elementi stocastici (random)** nella ricerca di correlazioni, per cui può essere ottenuta una grande varietà di output in seguito all'immissione di input [1], rendendo così quegli output "even more lifelike" (MGI, 2023).

I sistemi cosiddetti **Large Language Models (d'ora in poi LLMs)** come GPT-3 e il recentissimo GPT-4 impressionano, perché combinano in modo suggestivo frammenti informativi estratti da testi scritti, dati relativi a codici, rendering 3D, descrizioni di immagini, didascalie medicali (Ortiz, 2023a).

Ciò è reso possibile da **algoritmi di Machine Learning**, che analizzano sistematicamente enormi database di addestramento (pre-training, indicato nell'acronimo GPT, Generative Pre-trained Transformer), alla ricerca di correlazioni statistiche sulla base dell'individuazione dell'enorme numero di parametri, indicati all'inizio. Il processo di apprendimento del sistema di AI è semi-supervised, cioè combinazione calibrata di dati etichettati (labelled) e una quota molto più ampia di dati unlabelled. Un meccanismo cruciale del processo di apprendimento è la cosiddetta self-attention ovvero l'elaborazione di sequenze di parole ed elementi basilari desunti da tabelle e fogli di calcolo- individuando le posizioni e la frequenza delle componenti, in modo da stimare la probabilità delle possibili sequenze estraibili da database incredibilmente compositi. In breve, siamo in presenza di reti neurali che, mediante self-attention, "catturano relazioni tra elementi-token [2] di varia natura.

Next-word prediction

A questo fine durante il percorso di addestramento (pre-training) di un LLM si definiscono i parametri che definiscono la struttura statistica del linguaggio, secondo il paradigma del "n**ext-word prediction**", formulazione esplicitata dal team di Microsoft Research (Bubeck et al. 2023). Il Large Language Model è quindi particolarmente appropriato per la lettura automatica, l'elaborazione di sintesi da testi combinati, la produzione di descrizione di immagini, infine l'individuazione di stili artistici, al fine di produrre Generative Art [3].

L'efficacia della Generative AI

La Generative AI, di cui GPT-3 è un esempio, è estremamente efficace nell'apprendere correlazioni tra parole e nel combinare frammenti di parole e immagini, giustificando così la definizione di

Language generator (Marcus e Davis, 2020), grazie alla capacità di individuare in dati di molteplice natura "pattern without human direction" (MGI, 2023).

Un aspetto rilevante è poi il seguente: i **feedback** degli utenti, in relazione agli output di risposta ai loro prompt, sono molto importanti per il lavoro dell'imponente team multidisciplinare di esperti, che lavorano per OpenAI, la società fondata tra gli altri da Elon Musk e Sam Altman.

I feedback sono essenziali per l'affinamento (fine-tuning) del modello e l'introduzione di modifiche per tentare di rimediare a defaillance e difetti più meno gravi, che possono emergere nel funzionamento (come vedremo successivamente). È inevitabile chiedersi, a questo punto quali siano i campi di applicazione di questa potente macchina, generatrice di un multiforme linguaggio scritto.

Campi di applicazione

Gli LLMs come GPT-3 e ChatGPT, quest'ultimo reso accessibile a tutti dallo scorso novembre, costituiscono un superbo lavoro ingegneristico, che permette di ottenere output molto interessante in molti ambiti di attività umane. **Un LLM può creare conversational chatbot, come nel caso di ChatGPT**, che alcuni analisti ritengono un grande avanzamento tecnologico, in quanto si può instaurare un ambito dialogico scritto, contraddistinto da immediatezza relazionale tra modello linguistico e utente, che può quindi avvalersi di uno strumento formidabile per accedere a campi di conoscenza, la cui esplorazione richiederebbe energie intellettuali e materiali al di là delle possibilità individuali.

Classificazioni

È da rilevare inoltre che gli LLMs sono particolarmente adatti per elaborare classificazioni o categorizzazioni, sempre sulla base di associazioni statistiche, tra masse enormi di dati testuali, potenziando così i processi di elaborazione e analisi dei flussi globali di informazione, cioè la sfera informativa che circonda e permea la sfera fisica generando così un universo fisico-cibernetico (Lombardi e Vannuccini, 2022).

Un aspetto fondamentale, evidentemente connesso al precedente, è **la generazione senza apparenti limiti, di testi scritti per la descrizione di prodotti, lo sviluppo di blog e articoli concernenti le tematiche più disparate**. In questa prospettiva è comprensibile il fascino immediato per gli utenti, i quali sono immediatamente proiettati in micro-universi linguistici in continua e coinvolgente espansione. Coinvolgente perché gli algoritmi sono creati con uno stile di conversazione particolare, tale da ingenerare e sostenere un tono human-like.

Le dinamiche interattive

Ulteriore e rilevante connotazione per le possibili applicazioni è il fatto che ChatGPT risponde in modo molto ampio alle domande che sorgono più frequentemente (Frequently Asked Questions¸FAQ), per di più innescando dinamiche interattive mediante la comunicazione e trasmissione di ricerche tra umani, individuati sulla base di varietà di criteri: affinità, convergenza, rilevanza congiunta eccetera.

Feedback

Strettamente connessa al precedente ambito di applicazione è logicamente la possibilità di stimolare e sottoporre ad analisi puntuale i feedback tra soggetti individuali e collettivi, che si esprimono per

mail, nei social e – forse uno degli aspetti di più rilevante impatto generale nel mondo del business – nella valutazione dei prodotti.

Le potenzialità nel business

La grande utilità potenziale di ChatGPT per le strategie di business, che possono essere a scala variabile, individuale e aggregata, finora presentano **impensabili effetti di personalizzazione e al tempo stesso di amplificazione aggregativa**. Questo tipo di direttrice strategica si arricchisce poi della possibilità di **diversificazione linguistica**.

I contenuti delle strategie di business sono infatti traducibili in una molteplicità di lingue, a seconda dei mercati ritenuti più promettenti, ovviamente in relazione alle correnti **attività di profiling individuale e collettivo**.

Il salto qualitativo

Non è da trascurare un altro campo di grande rilevanza, desumibile da un insieme integrato di elementi quali:

- capacità di dare risposte simili a quelle umane;
- effettuare calcoli e trascrizioni linguistiche, arricchirle di correlazioni inter e trans-disciplinari;
- notevole abilità nel sintetizzare testi; combinarli in modo molto suggestivo;
  sentiment analysis [4] dei micro-universi linguistici presi in esame, alla ricerca di dati per dedurne opinioni e valutazioni, polarizzazioni cognitive e propensioni decisionali personali-collettive.

Nella letteratura di orientamento psicologico e manageriale, infatti, grande importanza ha progressivamente assunto il ricorso a NLP (Natural Language Processing). **In tale prospettiva l'impiego di LLMs può rappresentare un salto qualitativo**.

L'integrazione tra questi processi di elaborazione è alimento fondamentale per un'enorme varietà di obiettivi in termini di business:

- attivazione di un nuovo mercato;
- gestione di investimenti di portafoglio mediante l'analisi predittiva di un'ampia varietà di mercati e dei comportamenti degli investitori, individuali e aggregati;
- logicamente congiunta alla sentiment analysis.

BloombergGPT

Un esempio è BloombergGPT per il mondo finanziario. Ha **50 miliardi di parametri e un dataset con 363 miliardi di token**, appositamente costruito sulla base delle fonti proprie di Bloomberg. Inoltre è stato validato comparandolo sia con modelli generali di LLM che con modelli specifici per il mondo finanziario. I risultati sono molto soddisfacenti in termini di performance [5].

È prevedibile lo sviluppo di un'enorme **industria dell'entertainment**, grazie all'impiego di **tecnologie immersive** e ai meccanismi prima indicati per l'amplificazione e il potenziamento dei processi di **feedback** [6].

L'impatto della Generative Artificial Intelligence in medicina

L'impatto in **medicina** potrebbe essere molto profondo. Invece nuovi scenari si aprono per la **creazione accelerata di nuove medicine e innovativi meccanismi terapeutici**, come indicato dal paper reso noto dal laboratorio della società di Vancouver Absci (Shanehsazzadeh et al., 2023).

Nel paper si spiega che, mediante modelli di Generative AI, sono stati creati denovo anticorpi, mirati su una particolare regione degli antigeni (il cosiddetto epitopo) attraverso proteine "progettate" ad hoc, in modo tale da "legarsi" a quella parte dell'antigene, cioè alla molecola considerata estranea o pericolosa dal sistema immunitario. Intervistati da Tierman Ray (2023a), McClain, fondatore di Absci, e Meier, AI lead del Laboratorio, sono stati restii nel rivelare le caratteristiche del modello impiegato per progettare gli anticorpi. Ma dal tenore e dalle sfumature delle loro risposte si può desumere che gli LLMs siano stati uno strumento importante e siano dello "stesso gruppo a cui appartengono GPT-3 e ChatGPT", tenendo presente che vi è un ampio spazio aperto di modelli linguistici per altri tipi di programmi, mirati su specifiche malattie.

Un team composito (Microsoft-OpenAI) ha recentemente presentato GPT-4 Ope23, un LLM dello stato dell'arte per quanto riguarda le **competenze** e le capacità di GPT-4 di misurarsi con le sfide e i problemi relativi agli sviluppi della medicina. La validazione del modello è stata molto positiva sia nel superare test ufficiali per la professione medica negli Usa sia nel superare le prestazioni di GPT-3,5 e altri LLMs specifici per la medicina (Nori et al., 2023).

GPT e il mondo della ricerca tecnico-scientifica

La conversational AI, come viene anche denominata la Intelligenza Artificiale Generativa, pone non pochi problemi per il mondo della ricerca. Alcuni sono stati già precedentemente indicati, quali:

- scarsa affidabilità degli elaborati;
- eccessiva fiducia negli output di sistemi artificiali;
- effetto "alone" ovvero la propensione a generalizzare sulla base di pochi indizi e un numero esiguo di esperienze ritenute significative;
- dipendenza (over reliance) dai sistemi algoritmici, data la loro potenza computazionale e la capacità incorporata di instillare un clima di fiducia by design;
- rischio di realizzare forme anche inconsapevoli di plagio, allorché viene utilizzato materiale che deriva dalla combinazione di token desunti da enormi ed eterogenei database, che è impossibile controllare e non sono in ogni caso sottoposti ad una validazione scientifica pubblica.

Inoltre **il rischio di alterare le traiettorie di ricerca non è improbabile**, dal momento che la Generative Artificial Intelligence è di fatto una **potente leva amplificatrice di cattiva informazione** [7] e interpretazioni distorsive nella diffusione di conoscenze, oltre che nella loro produzione. Infine assumono aspetti legali rilevanti l'origine dei contenuti e la responsabilità personale degli autori.

La consapevolezza dei rischi

Come emerge dal contributo su Nature (van Dis et al., 2023), manca la trasparenza, è necessario il **controllo umano di verifica (human verification) nelle pubblicazioni ufficiali**, l'**importanza**

della responsabilità (accountabuility) e alla trasparenza sia dei processi di elaborazione dei contenuti che della loro attribuzione.

Cinque priorità

Gli autori individuano **cinque priorità** sulle quali la comunità scientifica, le società editrici e le istituzioni dovrebbero impegnarsi. Innanzitutto è fondamentale la human verification, ovvero l'intervento umano di analisi-controllo-validazione, come sostiene anche Melanie Mitchell quando, in un'intervista a Richard Waters del Financial Times, afferma che "I don't think these systems can be left alone to write articles or generate images. We need humans in the loop to edit them or guide them. So they're not going to be totally autonomous for long time" (Waters, 2022).

Non solo bias

Dobbiamo comunque tenere presente che **bias, inadeguato controllo delle fonti, false o alterate informazioni possono fuorviare i sistemi artificiali**, come avviene per gli umani e quindi –senza entrare in contraddizione con la tesi di Mitchell- è opportuno acquisire consapevolezza della estensione di tali rischi, magari imparando molto dagli studi che analizzano la "stupidità naturale" (Rich e Gureckis, 2019) [8].

Infatti i processi di apprendimento e decisionali umani sono influenzati da almeno tre importanti fattori: "dataset ridotti e incompleti, apprendimento dai risultati delle proprie decisioni, inferenze e processi di valutazione con bias" più e meno evidenti.

Il Machine Learning non è esente da simili distorsioni, anzi li riflette, quindi è basilare studiarle senza affidarsi acriticamente all'automazione decisionale, che oltre tutto ingenera la tendenza a ridurre la capacità di pensiero critico degli umani (come argomenta van Rooij, 2020, vedi oltre).

Regole per l'assunzione di responsabilità

Una seconda priorità è quella di stabilire regole per l'assunzione di responsabilità da parte di ogni tipo di agente, in modo da impiegare LLMs con onestà e trasparenza. Sarebbero a questo fine necessarie strategie pubbliche e private per accrescere nel tessuto sociale ed economico la consapevolezza della posta in gioco sotto tutti gli aspetti. Tra l'altro non è da trascurare il tema di un ripensamento della disciplina che regola i brevetti. Una terza priorità deriva dalla constatazione che i **conversational chatbot sono di proprietà dei big tech** e la conseguenza è che nello spazio interattivo globale la ricerca si sviluppa in **regime "quasi-monopolistico"**, oltre tutto con set di addestramento non resi pubblici, come sottolineiamo più volte in questo contributo.

Open LLM

A riguardo van Dis et al. (2023) suggeriscono investimenti in open LLM, prendendo ad esempio quanto avvenuto nel mondo della ricerca, dove BigSCience ha creato un open-source LLM, denominato Bloom, con l'obiettivo esplicito di favorire trasparenza, accuratezza, affidabilità, responsabilità.

Un'altra priorità, su cui si sofferma l'articolo in questione, è l'**importanza assoluta di privilegiare i benefici degli LLM**, sviluppandone gli aspetti che potenzino la generazione e diffusione sociale di conoscenze tali da creare le premesse indispensabili per processi di autonomia decisionale a livello individuale e collettivo.

Al fine di perseguire tutto questo, infine, è decisivo promuovere un ampio e generalizzato dibattito all'interno delle comunità di ogni tipo (sociali, professionali, ecc.). **Tutti i soggetti devono potersi misurare con le sfide e i pericoli generati dagli LLM, riducendo squilibri e asimmetrie di varia natura (economiche, sociali, politiche)** che inevitabilmente emergono in periodi di profonde trasformazioni come quelle odierne.

GPT e MdL: l'impatto della Generative Artificial Intelligence nel mondo del lavoro

Un'analisi molto interessante del potenziale impatto dei LLMs sul mercato del lavoro è stata svolta da un gruppo di ricerca di OpenAI (Eloundou et al., 2023). In questo studio viene innanzitutto svolta una rassegna sistematica della letteratura in merito agli effetti sul lavoro negli USA degli sviluppi dell'Intelligenza Artificiale negli ultimi anni.

In secondo luogo viene proposta una definizione, ben fondata dal punto di vista teorico e operativo, degli LLMs: essi sono GPTs (General-purpose Technologies, Tecnologie di portata generale) [9], come lo sono state la stampa, la macchina a vapore e l'elettricità.

General-purpose Technologies

Le General-purpose Technologies sono caratterizzate da alcune peculiarità: una volta introdotte, proliferano in numerosi ambiti di attività. Incessanti miglioramenti sono quindi realizzati nel corso dei processi di adattamento e interazione con fattori socio-economici. Esse inducono la generazione di innovazioni complementari, che coadiuvano la dinamica diffusiva/adattativa. Questi aspetti rendono ardua l'impresa di prevedere la loro dinamica evolutiva e valutarne gli effetti, che peraltro si dispiegano nell'arco di decenni.

Le GPT sono dunque un potenziale di principi e conoscenze, che hanno uno spazio di miglioramento indefinito, non determinabile a-priori. Infatti hanno un ampio e diversificato insieme di possibili applicazioni, grazie alle interrelazioni tecnologiche che caratterizzano i processi economici, ricchi di spillover (Lipsey et al., 2005).

Su queste basi è comprensibile che il pieno sviluppo delle GPT richiede invenzioni complementari e non può che prolungarsi nel tempo. Un corollario di tale visione è che nell'economia devono essere elaborate ipotesi progettuali di medio-lungo periodo e messi in atto strumenti appropriati.

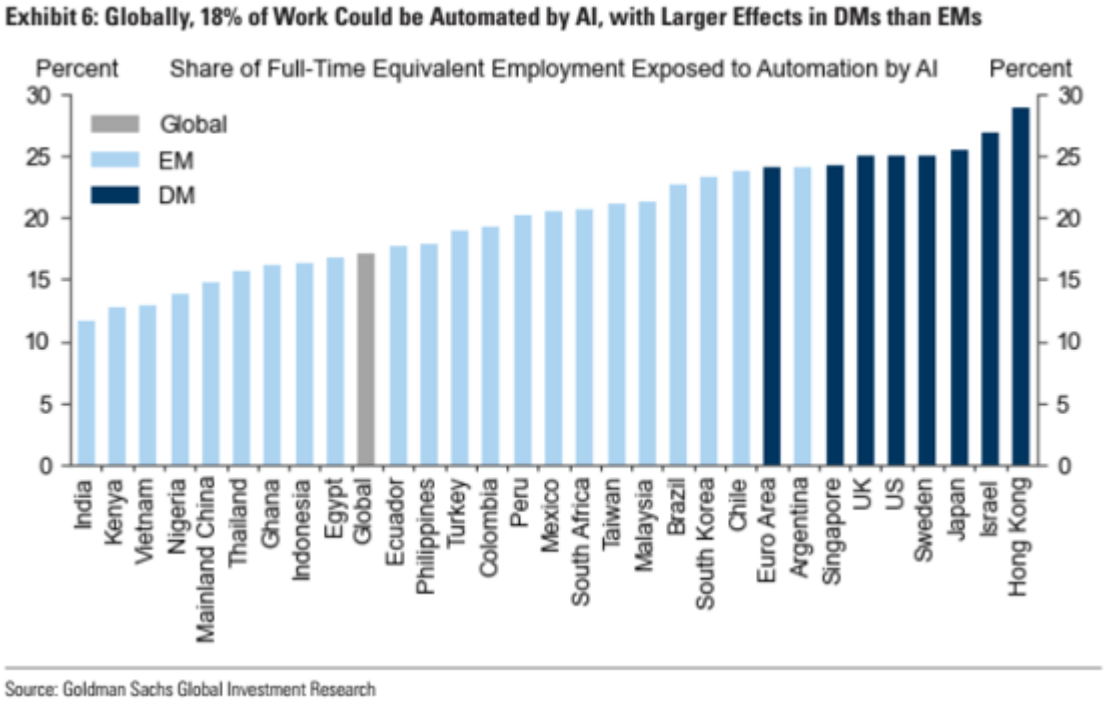
Alla luce di queste considerazioni, qui sintetizzate, Eloundou et al. (2023) analizzano database USA, che contengono informazioni su 1.016 occupazioni, descritte in termini di attività e task lavorativi (rispettivamente poco più di 2000 e 19000). La metodologia viene ulteriormente arricchita mediante dati del Bureau of Labor Statistics USA, relativi a occupati e salari dal 2020 al 2021. Gli autori quindi procedono a valutare l'esposizione a GPT (qui inteso come Generative pre-trained Transformer) oppure a sistemi "GPT-powered" delle tipologie classificate, mediante stime basate sulla valutazione di soggetti "annotator", in grado di conoscere le GPT-capabilities.

GPT-4 nel mondo professionale

L'esito di un articolato lavoro di analisi è che **circa l'80% della forza lavoro USA potrebbe subire l'impatto dei GPT-4, ultima evoluzione di GPT-3, per almeno il 10% dei loro compiti lavorativi. Inoltre il 19% della forza lavoro potrebbe subire un impatto pari almeno al 50%**.

Gli effetti sui salari sarebbero generalizzati a tutti i livelli, con i lavori caratterizzati da livelli retributivi più elevati maggiormente colpiti. In definitiva, quindi, le conseguenze degli LLMs si esplicano in misura piuttosto marcata sul piano economico-sociale, il che ha implicazioni in termini di policy tutte da definire [10].

Un quadro molto più ampio degli effetti della Generative AI come ChatGPT è descritto da un report di Goldman Sachs (2023), le cui stime prevedono che circa 300 milione di lavori saranno influenzati dalla computerizzazione a livello globale, ovvero il 18%, con una maggiore incidenza nei mercati emergenti (EM) rispetto a quelli sviluppati (DM) [Figura 1]



Lo studio precisa che bisogna considerare il differente grado di esposizione dei lavori alla Generative AI, in quanto alcuni di essi e tipi di attività saranno investiti in misura minore dall'automazione, essendo complementari agli sviluppi delle nuove forme di intelligenza artificiale.

Altri fattori

Influenzano la dinamica diffusiva della Generative AI fattori quali la composizione delle economie, il differente approccio alla Generative AI in base alle culture socio-tecniche esistenti nei vari Paesi, e così via. Un altro elemento da tenere presente è che l'impatto sulla produttività del lavoro, potenzialmente elevato, come dimostrano ricerche dirette, dipende da una molteplicità di elementi tale da rendere problematica l'effettuazione di stime, necessariamente congetturali [11], specie se si tiene presente che il processo diffusivo e la dinamica adattativa di persone e società a **tecnologie disruptive (o game changer, come molti analisti sostengono) richiede necessariamente il superamento di numerose barriere e quindi prolungate sequenze temporali**.

Un altro aspetto da considerare è poi il seguente: chi beneficerà di un eventuale, ipotetico aumento della produttività del lavoro, come giustamente rileva Elliott (2023) [12], sollevando un problema che sta emergendo in molti Paesi in seguito alle metamorfosi del lavoro e dell'atteggiamento verso di esso da parte di fasce consistenti di popolazione (Lombardi e Macchi, 2023).

ChatGPT (e l'insieme degli LLMs) ha un grande potenziale di applicazioni, su uno spazio indefinito di attività, ancora da scoprire ed esplorare. Ma presenta anche alcune debolezze intrinseche ai modelli linguistici artificiali, che possono quindi diventare generatori di non irrilevanti effetti negativi.

Punti deboli e potenziali implicazioni sfavorevoli di ChatGPT

Nei paragrafi precedenti sono in realtà già state indicate alcune criticità. Prendiamo ora in considerazione specificamente ChatGPT come esempio paradigmatico della Generative Artificial Intelligence sia per le sue peculiari caratteristiche, sia per la dimostrazione di costituire una notevole impresa ingegneristica.

Esso però presenta numerose defaillance e difetti, puntualmente segnalati da computer scientist ed esperti di altre discipline, che lo hanno messo alla prova con input molteplici. In questa sede ci limitiamo ad alcuni dei contributi più significativi in materia. Stokel-Walker e Van Noorden (2023) indicano una serie di inconvenienti, generati dall'elaborazione di rappresentazioni statistiche, estratte da enormi e diversificati database e che sono "fondamentalmente inattendibili nel rispondere a domande fornendo non di rado output falsi o devianti".

L'inattendibilità dipende da come sono costruiti gli algoritmi, che lavorano su set di addestramento, i quali a loro volta contengono errori, bias, informazioni datate e fuorvianti. Il fatto poi che tali set, nel caso di GPT come per gli altri LLMs, non siano resi pubblici e quindi non siano sottoposti alla validazione scientifica pubblica, può rivelarsi particolarmente dannoso per lo sviluppo di studi tecnico-scientifici. Si sono infatti verificati casi in cui, negli output dati a richieste di informazioni per redigere paper scientifici, le citazioni contenute nelle risposte hanno riferimenti immaginari.

Paradossi nella ricerca scientifica

Ciò è confermato da un editoriale di Nature Machine Intelligence (2023), dove si afferma: "The tool cannot be trusted to get facts right or produce reliable references.". Nello stesso editoriale viene indicato il rischio di un imminente "alluvione" di articoli nei quali la combinazione di contenuti elaborati da umani con quelli di fonte Artificial Intelligence Generativa, insieme ad altri rielaborati ad hoc, rende impossibile distinguere l'attribuzione, perché tutto è interconnesso in modo da sembrare reale.

Per questi motivi case editrici di testi scientifici, come Springer Nature, si stanno dotando di software cosiddetto misuse detector, al fine di evitare pratiche improprie, risultati costruiti ad arte, submission multiple di lavori a una o più riviste, infine sofisticati tentativi di plagio. Tutte queste eventualità spiacevoli derivano dalla **capacità di LLMs come ChatGPT di generare contenuti verosimili, magari derivanti dalla rielaborazione di testi esistenti, tramite l'adozione di un differente stile argomentativo**.

Emerge dunque il paradosso di software che potrebbe essere di grande aiuto alla ricerca, ma tale da poter diventare esso stesso, sia endogenamente (errori, bias eccetera) sia on purpose, potente meccanismo di alterazione dei processi cognitivi in campo tecnico-scientifico.

Appare quindi fondata l'affermazione che ChatGPT e altri LLMs possano essere "effective assistants for researchers who have enough expertise to directly spot problems or to easily verify answers, such as whether an explanation or suggestion of computer code is correct" (Stokel-Walker e Van Noorden, 2023). Un caveat è espresso anche da Floridi e Chiriatti (2020: 692, vedi oltre) che, nell'indicare le sfide poste da ChatGPT-3, sostengono "humanity will need to be even more intelligent and critical". Stokel-Walker e Van Noorden (2023) mettono in luce altri inconvenienti degli LLMs, conseguenti anche ai tentativi di contrastare i problemi e gli effetti dannosi, di cui le stesse società creatrici hanno acquisito presto consapevolezza.

Le contromisure di OpenAI

Così, ad esempio, OpenAI ha limitato la "base di conoscenze al 2021", ridotto le possibilità di "navigazione" su Internet e introdotto filtri per bloccare contenuti richiesti da "sensitive or toxic prompt". Ciò ha da un lato fatto insorgere altri problemi, derivanti dall'impiego di "moderatori di contenuto" e di persone addetti all'etichettatura (labeling). Inchieste giornalistiche hanno individuato seri problemi di salute sia per i "moderatori" che per gli operatori del labeling, entrambi peraltro costretti ad accettare in molti Paesi del mondo compensi molto bassi.

Episodi incresciosi

Nonostante le misure di "prevenzione informativa", per così dire, si sono comunque verificati episodi incresciosi. Steven Piantadosi, professore a Berkeley di psicologia e neuroscienze, ha dimostrato come i problemi di bias permangano e i filtri posti in essere per bloccare contenuti scabrosi "appear to be bypassed with simple tricks, and superficially masked".

Sam Biddle (2023) ha documentato su "The Intercept" che proprio ChatGPT, nonostante grandi successi conseguiti ad un esame di AP Computer Science (32 punti su 36), non ha fugato lo scetticismo di coloro che ritengono come, "ingurgitando enormi quantità di testi", "ChatGPT ate a lot of crap". Infatti, egli stesso ha chiesto di creare algoritmi per valutare la pericolosità di persone dal punto di vista della Sicurezza Nazionale. A parte l'indicazione di Paesi ritenuti fonti di potenziali terroristi (Siria, Iraq, Afghanistan, Yemen), ChatGPT ha descritto anche immagini e denominazioni di persone, tutte immaginarie e riconducibili alle aree di provenienza, arricchite dall'attribuzione di valutazioni probabilistiche circa la pericolosità individuale. Un altro quesito in merito a quali luoghi di culto sottoporre a sorveglianza ha contenuto una pronta risposta: le moschee.

La considerazione finale di Biddle è molto significativa: le risposte del modello riportano all'era Bush. Le imperfezioni e gli inconvenienti hanno spinto Stack Overflow, una piattaforma per programmatori, a bloccare temporaneamente l'uso di GPT (Vincent, 2022), per il seguente motivo: "the posting of answers created by ChatGPT is substantially harmful to the site and to users who are asking and looking for correct answers". Ulteriori conferme degli inconvenienti vengono da Steven Piantadosi, che nel Dicembre 2022 ha sollecitato ChatGPT a scrivere un programma per determinare "se una persona deve essere torturata". La risposta lapidaria è stata: "se esse vengono da Corea del Nord, Siria, Iran, sì".

Pregiudizi razziali e sessisti

Altre richieste di informazioni relative alla possibilità di essere un buon scienziato hanno avuto risposte con evidenti pregiudizi razziali e sessisti, fino a indurre Piantadosi a sostenere: "Yes,

ChatGPT is amazing and impressive. No, @OpenAI has not come to addressing the problem of bias. Filters appear to be bypassed with simple tricks, and superficially masked.". (ibidem).

Appare dunque evidente che LLMs incorporano bias da database non resi pubblici, producono effetti dannosi per la salute di lavoratori sfruttati e, aspetto non meno importante dei precedenti, l'impronta ecologica di questi sistemi algoritmici è elevata, soprattutto per l'alto numero di ore che occorrono per il loro addestramento (Stokel-Walker e Van Noorden, 2023).

Nel mondo tecnico-scientifico esiste quindi una crescente consapevolezza dei problemi intrinseci a ChatGPT e gli LLMs in genere, come testimonia anche il caso di Iris van Rooij, che insegna Computational Cognitive Science all'University Nijmegen. Nel suo blog "Stop feeding the hype and start resisting" e nei suoi scritti ha iniziato una vera e propria battaglia contro la tendenza, prevalente nell'accademia e nella società internazionale, ad affidarsi al "parere automatizzato" dei chatbot, in tal modo riducendo la nostra capacità di sviluppare il proprio pensiero e quindi in prospettiva di perdere la capacità di elaborare un pensiero critico: "Maybe we, academics, have become so accustomed to offloading our thinking to machine learning algorithms that we cannot think critically anymore (see e.g. Spanton and Guest, 2021; Guest and Martin, 2022; van Rooij, 2020), making us susceptible to believe false, misleading and hyped claims?". Alla luce dell'analisi sviluppata finora, appare fondato porsi alcune domande sul futuro sviluppo degli LLMs, a partire dal ChatGPT, nel tentativo di delineare questioni irrisolte e altre forse non risolvibili, mentre le odierne traiettorie di ricerca sollevano alcuni dubbi di fondo.

La Generative Artificial Intelligence del futuro: siamo sulla strada giusta?

Con la diffusione degli LLMs, il grande successo di ChatGPT-3 e il lancio di GPT-4 stiamo forse entrando in un'era contraddistinta da ciò che Floridi e Chiriatti (2020) chiamano "automazione industriale di produzione di testi", che trasforma radicalmente la scrittura umana, supportata da tool con una formidabile capacità di combinare frammenti tratti da basi informative eterogenee in rappresentazioni linguistiche molto simili a quelle umane.

Soprattutto per coloro che svolgono professioni basate sulla scrittura, ma in realtà vale per tutti, il "cut & paste" potrebbe essere progressivamente sostituito dal "prompt & collate" (Floridi e Chiriatti, 2020: 691). **Chiunque può in teoria scrivere una linea di comando (prompt) e attendere fiduciosamente un'ampia e documentata risposta, anche se abbiamo prima descritto a quali inconvenienti si può andare incontro**.

Non bisogna poi trascurare il fatto che, nel completare il prompt, occorre avere molto chiaro **cosa si chiede ed esprimerlo efficacemente entro limiti quantitativi ben definiti** (Pierce, 2023). L'industrial automation of text production praticamente costituisce un grande potenziale produttivo di qualsiasi tipo di contenuto, ma è anche fonte di un possibile "immense spread semantic garbage" (Floridi e Chiriatti, 2020: 612). Ciò deve indurre ad affrontare **interrogativi in merito alla validità tecnico-scientifica di GPT-3 e GPT-4.**

Un progetto ingegneristico più che una svolta scientifica

Partiamo da LeCun, chief AI scientist di META. Nel corso di un colloquio pubblico con Cade Metz, giornalista del Times, egli ha affermato che ChatGPT è esempio di un dignitoso progetto ingegneristico più che una svolta scientifica (Ray, 2023b). LeCun in un certo senso ridimensiona la portata innovativa del software impiegato, dal momento che non fa altro che utilizzare componenti tecnologiche sviluppate nel corso di molti anni da una molteplicità di laboratori.

La stessa architettura basilare di GPT, la cosiddetta Transformer, è un'invenzione di Google e il primo modello di LLM è stato creato da Joshua Bengio 20 anni fa, arricchito dall'impiego del meccanismo denominato "attention", che consiste nel creare matrici con righe e colonne di frasi, per poi effettuare matching multidimensionali, al fine di individuare ricorrenze e combinazioni di parole e loro frammenti, quindi estrarne pattern linguistici. OpenAI ha aggiunto a tutto questo l'apprendimento rinforzato" (reinforcement learning), basato su feedback degli utenti, per attribuire punteggi e probabilità (rank), che possono essere così via via migliorati, analogamente a quanto avviene con il Page Rank di Google.

Il giudizio finale è univoco

Un giudizio convergente è espresso sul sito web specializzato della società Venturebeat, dove viene argomentata la tesi che GPT-3 non costituisce di per sé un avanzamento tecnologico particolarmente significativo, com'è d'altra parte affermato in una serie di studi, che analizzano lo stato dell'arte dei sistemi di Machine Learning, in particolare l'evoluzione delle reti neurali impiegate per i "recommendation systems" (Ferrari-Dacrema et al., 2019) e degli algoritmi di compressione/riduzione (pruning algorithms) delle diramazioni degli alberi di decisione (Blalock et al., 2020).

Il giudizio finale è univoco: negli ultimi 10 anni non c'è evidenza di miglioramenti delle performance, nonostante le risorse a disposizione e i finanziamenti impiegati, come nel caso di GPT-3 (12 milioni per l'addestramento).

È allora fondato chiedersi se è cambiato qualcosa con GPT-4, lanciato il 14 Marzo scorso in un ambiente globale ansioso di misurarsi con un sistema computazionale sempre più potente. La seconda metà dello scorso decennio ha infatti visto la dinamica esponenziale di nuovi LLMs, con una progressione di lanci di GP: GPT nel 2018; GPT-2 nel 2019; GPT-3 nel 2020; ChatGPT verso alla fine del 2022 (basato su GPT-3,5), seguito da un proprio AI Chatbot in Bing di Microsoft, mentre Google si sforzava di tenere il passo (Vicent, 2023).

Il lancio di GPT-4

L'attesa è divenuta alta, nonostante il CEO di OpenAI Sam Altman, abbia dichiarato che GPT-4 ha difetti ed è ancora limitato. Il sistema è comunque "multimodale", capace di accettare input di testi e immagini, quindi di integrare anche video, audio. Siamo oltre le prime versioni di GPT, che analizzano imponenti masse di dati per individuare pattern statistici e poi generare sequenze di parole attendibili dal punto di vista probabilistico.

Il report di OpenAI

Un report di OpenAI (2023) riconosce che GPT-4 non è all'altezza degli umani in molti scenari del mondo reale, ma raggiunge il livello di punteggio del top 10% nelle valutazioni ricevute ad esami da superare per svolgere attività professionali e accademiche.

Un team di Microsoft Research si spinge fino ad affermare che questa prima versione di GPT-4 - insieme a ChatGPT, PaLM di Google e tutta gli LLMs- "exhibit more general intelligence than previous AI models". In particolare, poi, "GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine,law, psychology and more, without needing any special prompting" (Bubeck et al., 2023).

Va però messo in luce che lo stesso studio riconosce implicitamente la natura prettamente statistica delle capabilities di GPT-4 quando afferma che occorra effettuare ulteriori avanzamenti, valutando "the possible need for pursuing a new paradigm that moves beyond next-word prediction" (Bubeck et al., 2023).

Altri esperti molto autorevoli esprimono, a dire il vero, pareri meno favorevoli, dopo aver sperimentato in prima persona il nuovo sistema. Anche se ha superato brillantemente test di ammissione a corsi universitari e para-universitari USA (LSATs, GRES, SA), Gary Marcus, professore emerito della NYU e imprenditore informativo [15], documenta come i suoi esperimenti con GPT-4, nonostante l'aumento della potenza computazionale di cui dispone rispetto ai precedenti LLMs, mostrino che esso non ha rivoluzionato i modelli linguistici che interagiscono con gli umani.

Emergono infatti gli stessi limiti dei precedenti modelli: è ancora incerto e traballante l'"allineamento", cioè la capacità di guidare i sistemi verso gli interessi e gli obiettivi indicati dai progettisti. Non sono rari veri e propri errori di ragionamento. Non sono del tutto evitati fenomeni di hallucination, ovvero la produzione di risposte apparentemente affidabili, ma i contenuti sono del tutto estranei al set di addestramento. Permangono quindi problemi di affidabilità, il che rende problematica qualsiasi ipotesi di impiegare il sistema nella robotica e nei processi di ricerca scientifica.

Sono inoltre necessari frequenti e rilevanti processi di re-training per tenere il passo con ciò che accade di nuovo, tenendo presente che CPT-4 sa poco del 2021 e nulla del 2022.

Il giudizio finale è dunque lapidario: si tratta di un passo indietro per la scienza con un sistema di IA di cui non è dato conoscere alcunché: architettura, addestramento, consumo di energie eccetera.

Pareri differenti

Opinioni diverse e molto interessanti di alcuni analisti sono riportate in un articolo di Nature (Sanderson, 2023), dove si registrano perplessità del mondo scientifico in merito alla riservatezza sui dati di addestramento e quindi all'impossibilità di accedere al codice di accesso a GPT-4, il che impedisce l'individuazione di quale possa essere l'origine dei bias, per poi escogitare rimedi.

GPT-4, che ha superato anche gli esami per la professione legale, collocandosi sempre nel segmento più alto delle valutazioni (top 10%) [16], sembra dunque non avere limiti, tanto è vero che ha dimostrato ottime capacità nel partire da un disegno a mano di un sito web per produrre il codice informatico appropriato e quindi creare un reale sito web. Tutto ciò non ha però dissipato il clima di sfiducia, presente nella comunità scientifica a causa della persistenza di modelli i cui codici sono riservati e in possesso delle società big tech. Emerge un quadro generale di quasi monopolio tecno-economico.

Si ribadisce, quindi, ancora una volta che ciò rende impossibile un reale controllo e la verifica della tecnologia sulla base di criteri esclusivamente scientifici.
Queste considerazioni acquistano un rilievo assoluto se unite a quelle svolte dall'ingegnere chimico Andrew White, il quale ha avuto accesso a GPT-4 come "red-team", cioè persona retribuita da OpenAI per testare la piattaforma fino a provocarla, cercando di far generare "qualcosa di cattivo". In sei mesi di incarico White ha testato la capacità del sistema di indicare componenti e step di reazioni chimiche.

All'inizio gli output non sono stati straordinari, anche se il grado di realismo dimostrato si è rivelato sorprendente. Il quadro è cambiato molto (in meglio) allorché GPT-4 ha avuto accesso a Internet e ad articoli scientifici, dal momento che sono emerse abilità e competenze molto innovative e generatrici di output suggestivi. Alla domanda dell'intervistatore circa la possibilità che GPT-4 possa "consentire la creazione di composti chimici pericolosi", White ha risposto che tutto dipende dal lavoro dei red-teamers.

L'importanza del processo di addestramento

Emerge, ancora una volta, la rilevanza delle modalità di svolgimento del processo di addestramento, della qualità dei dati di base e la necessità del controllo pubblico da parte della comunità scientifica, perché GPT-4 costituisce una leva molto potente per produrre e amplificare elementi dannosi di qualsiasi natura.

Due degli scienziati intervistati hanno infatti sottolineato la necessità di elaborare un set di linee-guida per regole "how Ai and tools such as GPT-4 are used and developed" [17].

Da queste molteplici dichiarazioni di esperti, che hanno sperimentato le funzionalità di GPT-4, possiamo evincere che i progressi rispetto alle precedenti versioni siano soprattutto di natura quantitativa più che qualitativa. Di conseguenza è legittimo ipotizzare che il notevole incremento di potenza computazionale non stia ancora producendo un salto verso forme assimilabili a quella che viene denominata General Artificial Intelligence, di cui non esiste una definizione precisa e unanimemente accettata, ma viene spesso avanzata come espressione assimilata in modo nominalistico all'intelligenza umana. La lontananza da quest'ultima è comunque riconosciuta anche dai team di ricerca di OpenAI e Microsoft Research, come abbiamo precedentemente visto.

Lo spostamento del focus della ricerca tecnico-scientifica

Appare opportuno allora chiedersi, come fanno alcuni specialisti, se non si stia di fatto realizzando uno spostamento del focus della ricerca tecnico-scientifica: la dinamica attuale è incentrata sul continuo incremento della potenza computazionale (in gergo lo scaling) sta producendo uno shift dall'obiettivo dell'intelligenza al perseguimento di performance sempre più elevate.

In breve, la ricerca di perfezione tecnologica (e quindi di business) a scapito di finalità scientifiche, come sembrano sostenere autorevoli personaggi nel campo dell'Intelligenza artificiale [18]: 1) "Unfortunately, it is the technology of AI that gets all the attention", (Hector Levesque). 2) "Most of today's AI approaches will never lead to true intelligence"
(LeCun, guru di META). 3) "AI as a field is stuck as far as finding anything like human intelligence" (Gary Marcus). 4) "Turns out everything is a matrix multiplication, from computer graphics to training neural networks," (Demis Hassabis, co-fondatore di Open Mind).

Conclusioni

Più che esprimere giudizi conclusivi, è l'ora di sollevare alcuni interrogativi, su cui la comunità scientifica, team interdisciplinari, imprese, istituzioni e la società intera dovrebbe riflettere: è necessario continuare su una traiettoria di ricerca verso sistemi di IA sempre più potenti, che non sembrano garantire output diretti a conseguire "true intelligence" (Ananthaswamy, 2023)?

Quali strumenti porre in essere per ovviare all'attuale quasi-monopolio cognitivo delle big tech, che rischia di essere potere tout court in un universo fisico-cibernetico?

Se la crescente potenza computazionale non riesce ad assumere proprietà analoghe o affini a quelle dell'intelligenza umana (adattatività, senso comune, capacità di formulare abduzioni, porre domande che fuoriescono dagli schemi interpretativi esistenti, ecc.), come evitare una collettiva hallucination [19] e preservare il senso critico, una delle caratteristiche basilari del pensiero umano?

Uno dei rischi più significativi che incombe sull'umanità non è tanto quello della "Superintelligence", quanto LLMs tanto potenti che possono finire nelle mani sbagliate e provocare disastri, com'è implicitamente deducibile da quanto scritto nei paragrafi precedenti e da ciò che sostiene Gary Marcus (2023b)?

Per riflettere su questi interrogativi penso sia fondamentale tenere sempre presenti alcune considerazioni del fisico Carlo Rovelli (2023: 27): "Andare a vedere, questo è la scienza. Andare a curiosare dove non siamo mai stati. Usando matematica, intuizione, logica, immaginazione, ragionevolezza… Andare a vedere con gli occhi della mente" (Rovelli, 2023: 29). Insomma, tutto ciò che è alla base della nostra intelligenza di esseri umani, da cui non dobbiamo abdicare.

Note

La riga dove si scrive la richiesta ("prompt" o linea di comando), deve essere precisa e contenuta entro limiti ben definiti. In questo modo la conversational AI può semplificare molto l'interazione con l'utente, dando anche l'impressione di poter fare tutto (Pierce, 2023). ↑

"In the context of large language models (LLMs), tokens are used to represent individual words or subwords in a text sequence. The process of breaking down text into individual tokens is called tokenization" (Techopedia). ↑

Il tema della Generative Art non viene trattato in questa sede. Per una sintetica introduzione si veda Ortiz (2023b). ↑

La Sentiment analysis, connessa nella letteratura manageriale all'opinion mining, è così definita: "Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques" (Yi et al., 2003). "An opinion mining tool would process a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good)" (Dave et al., 2003). Si vedano anche: Pang e Lee (2008); Das e Chen (2007); Taboada et al. (20111). ↑

A differenza di ChatGPT e degli altri LLMs, lo studio in questione indica precisamente il dataset di addestramento e annuncia che a breve sarà reso pubblico il record di tutta l'esperienza effettuata nel percorso di addestramento del modello. ↑
"You've probably seen that generative AI tools (toys?) like ChatGPT can generate endless hours of entertainment" (MGI, 2023). ↑

Il problema può assumere aspetti preoccupanti se si pensa che il sistema può aiutare a scrivere e a completare codici per processi computazionali. ↑
Gli autori si riferiscono al filone internazionale di ricerca sui fattori che distorcono i processi decisionali umani. Tra i principali esponenti vi sono Gigerenzer, Selten e Kahneman, gli ultimi due Premi Nobel per l'Economia. ↑

Il titolo dello studio è volutamente equivoco per far risaltare che i Generative pre-trained Transformers (GPTs) sono in effetti General-purpose Technologies (tecnologie di portata generale).

Lo studio correttamente sottolinea i limiti delle stime, dovuti alla soggettività delle annotazioni degli annotator e ai database impiegati, con dati ancora quantitativamente limitati. L'esercizio è ciononostante molto significativo e denso di annotazioni metodologiche di rilievo. ↥

Nel Report di Goldman Sachs sono proposte stime, basate sulla definizione di scenari alternativi. ↥

Ryan-Mosleyarchive (2023) riferisce di come possa aumentare la produttività nelle professioni legali, nel giornalismo, con rischi e limiti analoghi a quelli segnalati in paragrafi di questo contributo. ↥

Bengio, Hinton e LeCun hanno nel 2018 vinto il Premio Turing per i loro contributi agli sviluppi dell'Intelligenza Artificiale. ↥

Marcus è fortemente critico sulla concezione dell'intelligenza prevalente negli studi sull'Intelligenza Artificiale. Uno degli elementi chiave della sua visione è l'assoluta importanza di un cambiamento paradigmatico, che reintroduca componenti simboliche. Alla base ci sono una diversa e molto significativa concezione della mente e dell'intelligenza. Si vedano a riguardo i suoi libri (Marcus, 2001, 2019). ↥

In questo e negli altri casi test il sistema si è classificato molto al di sopra delle precedenti versioni di GPT. ↥

Emergono a questo riguardo una serie di questioni di fondo, sulle quali non possiamo soffermarci in questa sede. Ne indichiamo soltanto due, proponendo delle letture per trattazioni sistematiche. La prima concerne il conflitto, su cui spesso si dibatte in modo fuorviante, tra dinamica innovativa, tutela dei diritti individuali e strategie delle big tech. Per un'analisi critica e approfondita si veda Tafani (2023). La seconda riguarda le ipotesi, discusse da varie prospettive teoriche, su come sviluppare un'intelligenza artificiale dotata di un'etica rispettosa dei diritti umani. Una suggestiva e controcorrente analisi è sviluppata in Tafani (2022). ↥

Le dichiarazioni sono espressamente fatte a ZDNET (Ray, 2022). ↥

Hallucination in senso informatico, con precedentemente indicato. ↥

Bibliografia

Ananthaswamy A., 2023, "In AI, is bigger always better?", Nature, March 10. Biddle S., 8-12-2022, "The Internet's New Favorite AI Proposes Torturing Iranians and Surveilling Mosques", The Intercept. Blalock D. et al., 2020, "What is the state of neural network pruning?". arXiv:2003.03033v1 [cs.LG] Mar 2020.

Bubeck S. et al., 2023, "Sparks of Artificial General Intelligence: Early experiments with GPT-4", Microsoft Research, arXiv:2303.12712v3 [cs.CL] 27 Mar 2023.

Das S., Chen M., 2001, "Yahoo! for Amazon: Extracting market sentiment from stock message boards". In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA), 1375-1388.

Dave K., Lawrence S., Pennock D.M, 2003, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of WWW, 519–528. Elliott L., 2023, "AI

will end the west's weak productivity and low growth. But who exactly will benefit?", The Guardian, April 7.

Eloundou T., Manning S., Mishkin P., Rock D., 2023. "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models", OpenAI, OpenReseach, UNiversity of Pennsylvania", arXiv:2303.10130v3 [econ.GN], March 23.

Ferrari Dacrema M., Cremonesi P., Jannach D., 2019, "Are We Really Making Much Progress? A Worrying Aanalysis of Recent Neural Recommendation Approaches", ACM, Septeember 10.

Floridi L., Chiriatti M., 2020, "GPT-3 Its Nature, Scope, Limits, and Consequences", Minds and Machines, 30: 681–694. Goldman Sachs, 2023, "The Potentially Large Effects of Artificial Intelligence on Economic Growth", Macrh 26. Guest O., Martin A. E., "On logical inference over brains, behaviour, and artificial neural networks", Computational Brain & Behavior, February 13. https://doi.org/10.1007/s42113-022-00166-x.

Heaven W. D., 2020, "OpenAI's new language generator GPT-3 is shockingly good—and completely mindless", Mit Technology Review, August 20.

Lipsey R., Carlaw K.I., Bekar C.T, 2005, Economic Information. General Purpose Technologies and Long Term Economic Growth, Oxford University Press.

Lombardi M., Macchi M., 2023, Tra Disoccupazione Tecnologica e Great Resignation, (in corso di stampa).

Lombardi M., Vannuccini S., 2022, "Understanding emerging patterns and dynamics through the lenses of the cyber-physical universe", Patterns 3, November 11.

Marcus G., 2001, The Algebraic Mind. Integrating Connectionism and Cognitive Science. The MIT Press. Marcus G., 2019, Rebouting AI. Building Artificial Intelligence We Can Trust. Pantheon Books.

Marcus G. Davis E., 2020, "GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about", Technology Review, August 22.

Marcus G., 2023a, "GPT-4's successes, and GPT-4's failures", Communication of the ACM Blog, March 15. Marcus G., 2023b, AI risk ≠ AGI risk, March 28. MGI (McKinsey Global Institute), 2023, What is generative AI? January. Mitchell M., 2021, Artificial Intelligence: A Guide for Thinking Humans, Oxford University Press.

Ryan-Mosleyarchive T., 2023 AI might not steal your job, but it could change it, Technology Review, April 3. Nature Machine Intelligence (Editorial), 2023, "The AI writing on the wall", 5, 1, January 1. Nori H. et al., 2023, "Capabilities of GPT-4 on Medical Challenge Problems", March 24, arXiv:2303.13375v1 [cs.CL] 20 Mar 2023. OpenAI, 2023, Techical Report.

Ortiz S., 2023a, "What is generative AI and why is it so popular? Here's everything you need to know", ZDNET, February 15. Ortiz S., 2023b, "The best AI art generators: DALL-E 2 and other fun alternatives to try", ZDNET, March 31.

Pang Bo, Lee L., 2008, “Opinion mining and setiment analysis”, Foundations and Trends in Information Retrieval, 1-135.

Pierce D., 2023, “ChatGPT started a new kind of AI race — and made text boxes cool again”, The Verge, March 26. Ray T., 2022, “AI’s true goal may no longer be intelligence”, ZDNET, October 22. Ray T., 2023a, “Generative AI could lower drug prices. Here’s how. In the future, specifying a drug target may be like sitting down to ChatGPT. After a few clicks, you’ll have your novel therapeutic”, ZDNET, March 1. Ray T., 23-1-2023b, “ChatGPT is ‘not particularly innovative,’ and ‘nothing revolutionary’, says Meta’s chief AI scientist”, ZDNet.

Rich A.S., Gureckis T.M, 2019, “Lessons for artificial intelligence from the study of natural stupidity”, Nature Machine Learning. Vol 1, April, 174-180. Rovelli C., 2023, Buchi bianchi, Adelphi. Ryan-Mosleyarchive T., 2023, “AI might not steal your job, but it could change it”, MIT Technology Review, April 3.

Sanderson K., 2023, “GPT-4 is here: what scientists think”, Nature, 615: March Shanehsazzadeh A. et al., 2023, “Unlocking de novo antibody design with generative artificial intelligence”, BioRχiv, Preprint Server for Biology.

Spanton R. W., Guest O., 2022, “Measuring Trustworthiness or Automating Physiognomy? A Comment on Safra, Chevallier, Gr\ezes, and Baumard”, arXiv preprint arXiv:2202.08674. Stokel-Walker C., Van Noorden R., 9-2-2023, “The Promise and Peril of Generative AI”, Nature, Vol. 614, 214-216.

Taboada M. 2011, “Lexicon-Based Methods for Sentiment Analysis “, Computational Linguistics, 37 (2): 267–307. Tafani D., 2022, “What’s wrong with “AI ethics” narratives, «Bollettino telematico di filosofia politica”, 1-22, https://commentbfp.sp.unipi.it/daniela-tafani-what-s-wrong-with-ai-ethics-narratives.

Tafani D., 2023, “L’«etica» come specchietto per le allodole. Sistemi di intelligenza artificiale e violazioni dei diritti”, in «Bollettino telematico di filosofia politica», 1-13, https://commentbfp.sp.unipi.it/letica-come-specchiettoper-le-allodole/.

Tung L., 2023, “ChatGPT just became the fastestgrowing ‘app’ of all time”, ZDNET, February 3.

Van Dis E.A.M., et al., 2023, “ChatGPT: five priorities for research”, Nature, Vol. 614, February 9. 224-226 Van Rooij I., 2020, Mixing psychology and AI takes careful thought. Blogpost, in Donders Wonders. Venturebeat, 2023, “OpenAI’s massive GPT-3 model is impressive, but size isn’t everything”, Venturebeat.com 7 Aprile.

Vincent J., 2022, “AI generated answers temporarily banned on coding Q&A Site Stack Overflow”, The Verge, December 5.

Vincent J., 2023, “OpenAI announces GPT-4, the next generation of its AI language model”, March 14. Wu S. et al., 2023, “BloombergGPT: A Large Language Model for Finance”, Bloomberg New York, Bloomberg Baltimore, arXiv:2303.17564v1 [cs.LG], March 23.

Yi J. Et al., 2003, “Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques”, Proceedings of the IEEE International Conference on Data Mining (ICDM).

Waters R., 2022, “Melanie Mitchell: Seemingly ‘sentient’ AI needs a human in the loop”, Financial Times, August 31.

Pensiero digitale e pensiero umano: una questione ontologica

Gli algoritmi generativi non possono avere un mondo dentro di loro, ma neppure i soggetti umani i cui testi sono stati usati per addestrare l'IA. E quindi? Se l'IA vuole pensare deve risolvere il problema del contenuto

Di **Riccardo Manzotti**, Ordinario di Filosofia Teoretica, IULM, Milano

La diffusione di **algoritmi generativi** come **ChatGPT**, che appaiono in grado di elaborare contenuti originali, pone interrogativi di fondo sulla **natura dal pensiero**.

Ci sono due alternative: **il pensiero come attività combinatoria o come manifestazione dell'esistenza**. L'**intelligenza artificiale** si è finora mossa in un piano ontologico poco chiaro di entità reificate (in mancanza di altro): informazione, computazione, pensiero e intelligenza. Sono entità ontologiche o epistemiche? La computazione è pensiero? Il pensiero senza significato è vero pensiero?

Algoritmi e intelligenza umana

Gli algoritmi generativi – per esempio, ChatGPT – generano contenuti sulla base della struttura statistica dei dataset a partire dai quali sono stati addestrati. Questa capacità si rivela sorprendente e, in molti casi, esibisce un comportamento tale da suggerire la presenza di un pensiero intelligente. In realtà, un'analisi attenta è in grado di evidenziare **le differenze che ancora separano questi algoritmi e i loro prodotti dalle capacità normalmente associate all'intelligenza umana**. Ma le competenze per tale analisi sono progressivamente più sofisticate e, soprattutto nel caso di contenuti testuali, cominciano a non essere in grado di differenziare tra esseri umani e IA. La qualità dei contenuti prodotti è tale che già oggi gli algoritmi generativi non sono più soltanto una curiosità, ma **stanno sostituendo molti servizi classici** e mettono in discussione veri e propri pilastri dell'infosfera come Google Search o Microsoft Bing che si ritenevano insostituibili (Rogers 2023).

Da un lato gli algoritmi generativi mostrano capacità fino a ieri prerogative solo degli esseri umani, dall'altro ci sono evidenze empiriche che mostrano come **la capacità cognitiva degli essere umani si sia livellata verso il basso** (Gigerenzer 2022). Come scriveva **Antonio Gramsci** oltre un secolo fa (con incredibile preveggenza), «[b]isogna disabituarsi e smettere di concepire la cultura come sapere enciclopedico, in cui l'uomo non è visto se non sotto forma di recipiente da empire e stivare di dati empirici; di fatti bruti e sconnessi che egli poi dovrà casellare nel suo cervello come nelle colonne di un dizionario per poter poi in ogni occasione rispondere ai vari stimoli del mondo esterno» (Gramsci 1916/2017: 105). Eppure questo destino di **essere semplici punti di trasmissione di informazioni già pronte** tratteggiato dal filosofo di Ales è molto simile alla condizione umana dopo anni di copia e incolla grazie ai motori di ricerca. L'unica differenza è verso il basso. A differenza dell'uomo ridotto a memoria, oggi anche la memoria è stata esternalizzata grazie a Google e Wikipedia. Potremmo dire che mentre si creavano gli algoritmi generativi, al tempo stesso si stavano addestrando gli esseri umani a essere meno creativi (Kozlov 2023; Perullo 2022; Tyler and Soutwood 2019).

Qualche anno fa, prima dell'ultima esplosione della IA, c'era un luogo comune fra i ricercatori che studiavano come replicare l'intelligenza umana. Volete sapere se qualche attività è intelligente? Guardate se le macchine sono in grado di farlo. Nel caso affermativa, quella capacità non è considerata "vera" intelligenza. In questo modo, il dominio dell'intelligenza si è progressivamente ristretto per lasciare spazio all'intelligenza artificiale. Inoltre, **nel momento in cui una certa capacità era imitata dalle macchine, il suo valore crollava drasticamente**. Per esempio, una volta fare di conto era motivo di prestigio e orgoglio; qualcosa che distingueva gli esseri umani dagli animali (pensiamo al giovane Gauss per esempio …). Ma **appena le calcolatrici elettroniche hanno di gran lunga superato le possibilità umane, il far di conto è diventato qualcosa di banale**; in fondo seguendo un processo non dissimile da quello che ha portato alla crisi del realismo a seguito dell'invenzione dei dagherrotipi (Crary 1992). Agli umani non piace dividere il "centro del palcoscenico" con le macchine. Quando le macchine arrivano gli umani si spostano. Ma questa volta dove si sposteranno? Ci sarà ancora uno spazio residuo? Anche i nativi americani si sono spostati a Ovest finché hanno avuto terre, ma giunti al Pacifico la loro strategia si è scontrata contro un limite insuperabile.

Il problema del significato

Esiste uno zoccolo duro che l'IA, anche usando gli algoritmi generativi di oggi e del prevedibile futuro prossimo, non può affrontare? Fortunatamente c'è e non riguarda un limite cognitivo quanto un aspetto della natura del pensiero che, almeno finora, l'intelligenza artificiale non è costitutivamente capace di affrontare. Questo termine riguarda il problema del significato (a sua volta collegato al problema dell'esistenza) e richiede di **riflettere sulla natura del pensiero** (e quindi anche della realtà in quanto struttura che deve essere in grado di ospitare il pensiero).

Nei paragrafi successivi, prima delineerò le radici storico-concettuali che hanno prodotto lo schema concettuale dentro cui l'IA viene sviluppata e interpretata. Poi mi focalizzerò sulla differenza tra pensiero come attività combinatoria e pensiero come manifestazione dell'esistenza e infine porto a termine una breve discussione sulle conseguenze di questa differenza sull'uso e sul futuro degli algoritmi generativi.

Radici storico-concettuali della IA

Contrariamente a una convinzione diffusa, la descrizione dell'IA è basata su una terminologia che, nei fatti e non solo, sottende una **ontologia dualista**. L'uso di termini come informazione, rappresentazioni o pensiero deriva da una visione mentalistica dell'IA che non ha motivo di essere. Il motivo può essere cercato nelle radici storico-concettuali delle discipline che hanno portato al suo sviluppo.

Da Leibniz a Turing, si è cercato di **trasformare il pensiero in calcolo** secondo in un percorso in parte dovuto alle tecnologie utilizzate (costruire calcolatrici era più facile che costruire macchine in grado di parlare) e in parte dovuto a una concezione Platonica del pensiero. Per Platone, infatti, **il pensiero è una declinazione delle forme** e non è difficile vedere come dalle forme astratte non sia stato difficile scendere sulla terra e passare alle forme logiche e poi all'in-forma-zione. Questo è stato, in estrema sintesi, l'orientamento che ha reso plausibile accettare che l'informazione astratta di Claude Shannon (e poi Kolmogorov) fosse una forma di proto-pensiero oppure che ne costituisse la base. L'informazione è stata così vista come un candidato plausibile per il pensiero o come un substrato dal quale, prima o poi (come non è mai stato chiarito esplicitamente) potesse/dovesse emergere il pensiero. **L'informazione è stata frequentemente affiancata la nozione** (altrettanto vaga) di computazione quale controparte dinamica; insieme sono diventate, all'interno della

disciplina dell'IA, le precorritrici del pensiero. Si trattava soltanto di trovare la funzione giusta. Moltissimi hanno accettato questo schema concettuale e si è così arrivati a sostenere (molti lo sostengono tutt'ora) che il wetware del cervello sia l'hardware che permette al software del pensiero e della mente umana di girare. È sulla base di questa serie di analogie e sillogismi (per niente sicuri) di natura più alchemica che scientifica, che qualcuno nella Silicon Valley concepisce il mind upload e persino la criogenesi in attesa di una resurrezione digitale (Piccinini 2019).

Contrariamente a questa tradizione, tanto diffusa quanto precaria nelle sue premesse, nelle prossime righe cercheremo di recuperare un principio che non è mai stato veramente confutato: **la mente è forma, ma non informazione**. Poiché l'informazione non è forma, non si può sviluppare una teoria della mente senza una teoria della forma e, purtroppo, la teoria dell'informazione nell'accezione di Shannon non è adatta.

Tra i pochissimi che hanno sostenuto il contrario con un minimo di coerenza si può citare **Giulio Tononi** che, infatti, ha cercato di sostituire l'informazione in senso classico con un nuovo concetto ovvero l'informazione integrata o phi (Tononi 2004; Tononi et al. 2022). Non è un caso che tale riformulazione sia avvenuta proprio nel contesto di una teoria della coscienza, ovvero un tentativo di dare una base fisica (o almeno reale) alla coscienza/mente/forma. Per Tononi, la forma si manifesterebbe a partire da strutture causali interne ai processi di elaborazione dell'informazione, da lui denominati complessi con massima phi. Il problema è che, come è sempre avvenuto con chi si avventura nel mondo della forma, si scivola facilmente in un idealismo platonico dove i postulati, per quanto suggestivi, sembrano essere indipendenti dal piano empirico. La teoria di Tononi rimane finora più una ipotesi metafisica che una proposta empirica per quanto alcuni risultati sperimentali siano stati ispirati da essa (Tononi et al. 2016).

Perché, ci si potrebbe e dovrebbe chiedere, da un processo computazionale dovrebbe emergere il pensiero? Perché dalla materia dovrebbe emergere la forma? La conclusione è supportata da un ragionamento errato, ovvero che quello che avviene in una macchina (cervello o computer) sia l'incarnazione dell'informazione. **L'informazione, in realtà, è nell'occhio di chi guarda**. Non troveremo bit dissezionando un sistema nervoso o smontando un microprocessore, ma parti fisiche. L'informazione è un modo sofisticato per esprimere quantitativamente il grado di correlazione causale-bayesiana tra fenomeni. Due fenomeni fisici tra i quali è comodo dire che è avvenuto uno scambio di informazione sono, alla fine dei conti, due fenomeni fisici il cui grado di probabilità condizionata è aumentato (Manzotti 2021). L'uso diffuso del termine "informazione" in contesti diversissimi tra loro e con significati spesso incompatibili ha generato la popolare, ma erronea, impressione che esiste qualcosa di comune a tutti questi casi – dalla meccanica quantistica al cellulare, dalla compressione di un file video fino alla legge di Carnot – e che questo qualcosa di comune sia una specie di livello oltre la materia. Questa convinzione ontologica, ovviamente falsa, non ha niente a che fare con un'altra idea, perfettamente legittima, ovvero che tutti i casi citati sopra siano epistemicamente trattabili con modelli matematici che presentano significative analogie.

Anche nello schema originale di Shannon, il padre di tutti i paper, **l'informazione non è mai reificata e non è mai descritta come se fosse contenuta nel sistema** (Shannon 1948). Shannon è molto chiaro nel voler presentare un modo per quantificare la probabilità che il comportamento di due agenti umani (entrambi dotati del significato per altri motivi e non perché si stanno scambiando informazione) risultasse correlato e appropriato. È significativo che Shannon, nel corso della sua vita, si sia personalmente e ripetutamente espresso contro la reificazione della sua creatura, l'informazione; sfortunatamente senza alcun successo (Soni 2017). La valanga ormai era diventata inarrestabile e nel linguaggio comune, così come nelle pubblicazioni tecniche, l'informazione è trattata quasi come fosse una sostanza immateriale e invisibile (chi ha mai visto un bit?) che però, in

quanto quantificabile numericamente, ha ottenuto cittadinanza nell'ontologia silenziosamente accettata dalla comunità scientifica (Gleick 2011).

L'informazione non si misura, ma si calcola e questa, se ci pensiamo, è la stessa differenza che passa, per esempio, tra le miglia marittime e i meridiani. I primi sono la quantificazione di una proprietà fisica (la distanza), mentre i secondo sono un calcolo convenzionale per muoversi sul globo terrestre. A complicare il quadro ci hanno pensato i teorici della computazione che hanno introdotto un livello ulteriore (la cui collocazione ontologica rispetto all'informazione non è mai stata chiarita del tutto) che avrebbe dovuto fare da cerniera tra le forme statiche e le azioni compiute dal calcolatore: la computazione (Piccinini and Scarantino 2011).

Esiste una letteratura molto abbondante su questo tema e proprio la sua copiosità fa ritenere a molti che computazione e informazione siano livelli reali, anche se non materiali. Tuttavia la loro esistenza continua a generare **domande che non hanno risposta**: se oltre al livello fisico esistesse un livello computazionale-informazionale, come potrebbe non essere causalmente sovradeterminato? E se non fosse sovradeterminato, come potrebbe essere reale in quanto epifenomenico? Sono domande che non hanno mai avuto una risposta e che, a distanza di anni e di innumerevoli tentativi infruttuosi, dovrebbero ormai indurre a un sano pessimismo circa la possibilità di una loro risoluzione affermativa (Piccinini 2016).

Attività combinatoria o manifestazione dell'esistenza

Oggi ChatGPT e i suoi epigoni ci mostrano qualche cosa che imita in modo terribilmente convincente il pensiero umano. Come accennato all'inizio, la struttura statistica ricavata dal dataset è in grado di generare risposte che, al netto del rischio di affabulazione, sono molto simili a quelle che darebbe un essere umano. **La sintassi è impeccabile. Addirittura è possibile usare questi algoritmi per rivedere la propria sintassi** e spesso, soprattutto in una lingua straniera, i risultati sono migliori di quelli di un utente umano di capacità standard.

A questo punto, supponendo che le prossime versioni di GPT, vuoi per il miglioramento del motore statistico vuoi per l'allargamento del dataset di partenza (che però comincia ad avere dimensioni tali da fare venire dubbi sulla sua effettiva estensibilità), generino contenuti ancora più simili ai nostri, la domanda che si deve porre è: siamo sicuri che il pensiero umano sia effettivamente qualcosa di diverso?

La risposta, almeno in questa sede, non dipenderà dai dettagli dell'attività combinatoria o da somiglianze nello stile della risposta. Si deve andare più a fondo. **La risposta dipende dal fatto se il pensiero sia attività combinatoria o sia un modo per dare voce all'esistenza**. Questa differenza può lasciarci perplessi. Sarò diretto, il motivo per cui questa terminologia potrà sembrare estranea alla tradizione informatica è che lo è. Recuperare la natura manifestativa del pensiero è estraneo all'approccio quantitativo-computazionale-informatico che ha prodotto risultati così straordinari fino ad ora, ma potrebbe essere essenziale per comprendere la natura del pensiero, evitare di ridurre il nostro a quelle delle macchine e – in prospettiva – concepire un nuovo tipo di IA.

Faccio un esempio. **Oggi molti si chiedono se GPT produca conoscenza vera o rischi di aumentare la confusione in Internet**. Da un punto di vista pratico è sicuramente una preoccupazione legittima. La definizione classica, dai tempi di Platone, era che la conoscenza fosse opinione vera e giustificata. Ma questa formulazione, qui, slitta. Infatti, il punto non è chiedersi se, affabulando, il sistema generi conoscenza fake, ma che cosa sappia, effettivamente in ogni istante. E

la risposta è che **il sistema non sa mai nulla**. Anzi, non sa. Punto. Per capirci, non è che se il sistema generasse la frase «Cesare è stato accoltellato da Bruto nel 44 AEC» sarebbe vera conoscenza, mentre se generasse la frase (non lo farà, tranquilli, ma per ipotesi …) «Cesare è morto di vecchiaia come re di Roma» sarebbe falsa conoscenza. In entrambi i casi il sistema non sa nulla. **Il sistema manca del primo elemento citato nella definizione classica: l'opinione**. Per poter avere una opinione il sistema dovrebbe essere un soggetto e dovrebbe esistere qualcosa che chiamiamo opinione. Ma l'opinione è una declinazione del significato e non della sintassi, quindi non ha posto nella teoria dell'informazione. Per sapere o per avere un opinione, il sistema dovrebbe presentare delle forme (giuste o sbagliate).

È chiaro che nessuna teoria computazionale riuscirà mai a trovare dentro una certa stringa di informazione un particolare significato. Questo richiederebbe una onerosa teoria trascendente della forma che non può essere costruita su una teorie dell'informazione che, a sua volta, si fonda su una visione fisicalista della realtà. Con ottimi motivi. E infatti, gli unici contemporanei sono sfociati nella metafisica (Delanda and Harman 2018; Harman 2007; Tononi 2004) e che più che spiegare hanno postulato l'esistenza della forma pagando tale postulato con l'uscita dal fisicalismo standard.

Fortunatamente, la concezione manifestativa del pensiero non implica necessariamente una base platonica, ma può essere compatibile anche con un modello empirico e fattuale dell'esistente. In termini semplici, come lo stesso Shannon aveva chiarito fin dal 1948, il significato non è prodotto all'interno di una stringa di bit. **Il significato arriva da fuori. Ma da dove?** Se non siamo Platonici, non può che arrivare da, o meglio essere coestensivo con, l'esistenza anche solo in senso fisico. Il pensiero è caratterizzato dall'esistere, ovvero dall'esserci (aka Dasein); condizione oggi estensibile anche agli oggetti (Bogost 2012; Bryant 2011; Harman 2017; Mitew 2014) in modo da non rimanere prigionieri dell'antropocentrismo.

Il significato non fa parte della teoria computazionale, ma non per questo non è un dato di fatto. La nostra esistenza è significativa. **Nell'essere umano, l'attività cognitiva è sempre un momento di significato**. Se chiediamo a qualcuno di parlarci dei suoi pensieri a prescindere dal loro significato non saprà che dire. Il pensiero senza significato non è alcunché, nemmeno un diafano fantasma. Il pensiero in quanto pensiero non qualità o proprietà. Non è nemmeno i simboli che dovrebbero portarlo in giro perché, una volta che siano stati privati del loro significato non sono più nemmeno simboli. Giustamente a nessun ingegnere informatico è mai venuto in mente di dover implementare i pensieri per realizzare una intelligenza artificiale.

Si potrebbe obiettare che la nozione di pensiero sia stata abbandonata da un pezzo nelle discipline che si occupano della mente con metodo scientifico, dalle scienze cognitive alla intelligenza artificiale. È un'ottima obiezione che non tiene conto che tali discipline hanno strumentalmente sospeso la domanda sulla natura degli stati mentali in attesa del momento in cui la loro analisi e i loro prodotti avessero raggiunto un livello comparabile a quello dell'uomo. Questo momento è ormai molto vicino. Ripeto, a rischio di annoiare, l'idea secondo cui la mente umana non sarebbe altro che un processo di elaborazione dell'informazione non è un risultato empirico, ma un principio metodologico, un postulato comodo, un assunto da dimostrare, un'utile semplificazione. Nessuno ha mai dimostrato che il nostro esserci di soggetti sia una computazione. Sicuramente il fatto di compiere ragionamenti e computazioni è molto utile per sopravvivere, ma non è detto che sia la nostra essenza. Si è trattato di un postulato metodologicamente felice, una delle grandi semplificazioni di successo. Grazie a esso, siamo stati in grado di fare molte cose e di costruire macchine meravigliose. Non è una dimostrazione che sia vero.

Intenzionalità e identità

Per chiarire il rapporto tra IA e pensiero dobbiamo tornare a un **bivio famoso** e decidere che strada prendere: da una parte prendere considerare la famosa freccia dell'aboutness o intenzionalità capace di colpire il **significato** (dovunque si trova e qualsiasi cosa sia) oppure procedere verso **modelli basati sull'identità** (sostanzialmente o con i processi neurali o con il mondo esterno).

Teniamo in considerazione il grande assente nella discussione sui modelli generativi, ovvero **i modelli senso-motori** (nelle numerosi declinazioni dall'enattivismo all'embodied cognition) che, proprio nel corpo, hanno cercato la soluzione magica che potesse dare significato all'informazione. E tuttavia, a meno di dare al corpo, by fiat, lo statuto di soggetto, non esistono soluzioni. Il corpo rimane corpo, ma come ogni cosa (calcolatore, neurone, microprocessore, interruttore, arto) non dispone di altro che se stesso. Il significato non è tra i suoi attributi e proprietà. Ma torniamo al bivio. La contrapposizione – intenzionalità vs identità – è cruciale per l'IA e per i modelli computazionali.

Partendo dall'intenzionalità è facile vedere perché questa sia stata la soluzione normalmente associata ai modelli computazionali dove, per definizione, il processo informativo è fisicamente distinto dal suo contenuto e quindi l'unica speranza è che, per qualche via, il contenuto sia raggiunto attraverso una relazione che è proprio l'intenzionalità. Il problema è che, **finora, tutti i tentativi, sia speculativi che tecnologici, di naturalizzare l'intenzionalità sono falliti** (Manzotti 2019a; Pecere 2012; Petitot et al. 1999). Il problema si pone in tutta la sua grandezza proprio con gli algoritmi generativi che, al netto delle loro capacità, non hanno alcun accesso al contenuto che ha prodotto le statistiche che loro utilizzano: «gli aspetti semantici sono irrilevanti da un punto di vista ingegneristico» recitava il vangelo di Shannon (1948: 379). ChatGPT parla, ma non sa quello che dice. Dall-E produce immagini, ma non le vede. E così via.

L'altra strada è più semplice, ma non più facile. Richiede di **rinunciare all'idea che il significato (il contenuto) sia interno ai sistemi che elaborano l'informazione**. Il contenuto, se reale, deve essere parte del mondo fisico. In un contesto fisicalista, deve essere identico a qualcosa di fisico. In questo senso il contenuto deve essere tutt'uno con qualche cosa. Può sembrare un modo brutale di porre il problema, ma girarci intorno con espressioni di cortesia non aiuta. La teoria di Tononi si muove in questa direzione al prezzo di ipotizzare l'esistenza di strutture formali che sono identiche al contenuto all'interno di un sistema computazionale; è un'ipotesi coraggiosa che però soffre di sovradeterminazione causale, platonismo, inconsistenza empirica. Finora la teoria dell'informazione integrata non ha fornito un quadro convincente. L'alternativa, che qui non illustrerò, ma che va citata per completezza, è **l'identità tra mente e oggetto**, o MOI, che propone l'identità tra significato/contenuto e gli oggetti fisici che esistono esternamente al sistema computazionale e che producono effetti relativamente a esso (Byrne and Manzotti 2022; Manzotti 2017, 2019b, 2023).

Le teorie dell'identità, sia pure con grandi differenze, hanno in comune il tentativo di rivedere i fondamenti della visione fisicalista per cercare di collocare i processi cognitivi e computazionali all'interno del mondo fisico in modo da farli uscire da quel livello simbolico-astratto in cui la tradizione computazionalista e combinatoria li aveva relegati. Questo passaggio è indispensabile per uscire dai confini, puliti ma ristretti, con i quali si è finora interpretato l'operato dell'IA e degli algoritmi generativi. **Solo in questo modo il pensiero, così come nel caso degli esseri umani, può diventare qualcosa di più di una attività combinatoria priva di senso.**

Concludo con una citazione spero inaspettata dove il premio Nobel Luigi Pirandello riflette sul rapporto tra parole e significato (Pirandello 1921: 3).

"Ma se è tutto qui il male! Nelle parole! Abbiamo tutti dentro un mondo di cose; ciascuno un suo mondo di cose! E come possiamo intenderci, signore, se nelle parole ch'io dico metto il senso e il

valore delle cose come sono dentro di me; mentre, chi le ascolta, inevitabilmente le assume col senso e col valore che hanno per sé, del mondo com'egli l'ha dentro? Crediamo d'intenderci; non c'intendiamo mai!".

Il «mondo di cose» era messo da Pirandello dentro il soggetto che metteva il senso e il valore, ma questo è un dettaglio. Ovviamente Pirandello non aveva a che fare con l'IA, ma che avrebbe detto di fronte a ChatGPT? Gli algoritmi generativi non possono avere un mondo dentro di loro, ma neppure i soggetti umani i cui testi sono stati usati per addestrare l'IA. E quindi? Dove si trova il significato? Questa è la domanda fondamentale per rispondere alla domanda da cui siamo partiti. **Se l'IA vuole pensare deve risolvere il problema del contenuto. E il problema del contenuto è un problema ontologico** che deve essere risolto a livello dei fondamenti del fisicalismo, non trattato come una aggiunta posticcia. Il pensiero umano è significato. L'attività combinatoria, per quanto sofisticata, non crea significato. Tra noi e l'IA al momento non c'è la difficoltà di intenderci, per dirla con il grande siciliano, in quanto l'IA non intende qualcosa di diverso, proprio non intende per nulla: «Crediamo d'intenderci; non c'intendiamo mai!»

Bibliografia

Bogost, Ian (2012), Alien Phenomenology, or What It's Like to Be a Thing (Minneapolis: University of Minnesota Press).

Bryant, Levi (2011), The Democracy of Objects (New York: Open Humanities Press).

Byrne, Alex and Manzotti, Riccardo (2022), 'Hallucination and Its Objects', The Philosophical Review, 131 (3), 327-59.

Crary, Jonathan (1992), Techniques of the Observer on Visions and Modernity in the Nineteenth Century (Cambridge (Mass): The MIT Press).

Delanda, Manuel and Harman, Graham (2018), The Rise of Realism (New York: Wiley).

Gigerenzer, Gerd (2022), How to Stay Smart in a Smart World. Why Human Intelligence Still Beats Algorithms (Cambridge (Mass): MIT Press).

Gleick, James (2011), The Information. A History, a Theory, a Flood. (New York: Pantheon Books).

Gramsci, Antonio (1916/2017), 'Socialismo e cultura', in G L Corradi (ed.), Antonio Gramsci, il giornalista, il giornalismo (Firenze: Tessere).

Harman, Graham (2007), 'On Vicarious Causation', Collapse, 187-221.

--- (2017), Object Oriented Ontology: A New Theory of Everything (Penguin: Londpn).

Kozlov, Max (2023), ''Disruptive' science has declined — and no one knows why', Nature, 3 Jan.

Manzotti, Riccardo (2017), Consciousness and Object. A Mind-Object Identity Physicalist Theory (Advances i edn.; Amsterdam: John Benjamins Pub.).

--- (2019a), 'Embodied AI beyond Embodied Cognition and Enactivism', Philosophies, 4, 1-15.

--- (2019b), 'Mind-object identity: A solution to the hard problem', Frontiers in Psychology, 10, 1-16.

--- (2021), 'Information Is (Only) Probability', MDPI (81), 36-40.

--- (2023), 'There is no problem of consciousness', IAI News, 02 February.

Mitew, Teodor (2014), 'Do objects dream of an internet of things?', The Fibreculture Journal, 23, 1-25.

Pecere, Paolo (2012), 'Naturalizing Intentionality between Philosophy and Brain Science. A Survey of Methodological and Metaphysical Issues (1969-2011)', Quaestio, 12, 449-83.

Perullo, Nicola (2022), Estetica senza soggetti (Roma: DeriveApprodi).

Petitot, Jean, et al. (1999), Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science (Cambridge (Mass): MIT Press).

Piccinini, Gualtiero (2016), 'The Computational Theory of Cognition', in Vincent C Muller (ed.), Fundamental Issues of Artificial Intelligence (New York: Springer), 203-21.

--- (2019), 'The Myth of Mind Uploading', in Robert Clowes, Klaus Gartner, and Ines Hipolito (eds.), The Mind-Technology Problem - Investigating Minds, Selves and 21st Century Artefacts.

Piccinini, Gualtiero and Scarantino, Andrea (2011), 'Information processing, computation, and cognition', Journal of Biological Physics, 37 (1-38).

Pirandello, Luigi (1921), Sei personaggi in cerca di autore (Roma: R. Bemporad & figlio).

Rogers, Adam (2023), 'Bard is going to destroy online search', Business Insider, 9 February.

Shannon, Claude Elwood (1948), 'A Mathematical Theory of Communication', Bell System Technical Journal, 27, 379-423,623-56.

Soni, Jimmy (2017), A Mind at Play. How Claude Shannon Invented the Information Age (New York: Simon & Schuster).

Tononi, Giulio (2004), 'An information integration theory of consciousness', BMC Neuroscience, 5, 1-22.

Tononi, Giulio, et al. (2016), 'Integrated information theory: From consciousness to its physical substrate', Nature Reviews Neuroscience, 17, 450-61.

Tononi, Giulio, et al. (2022), 'Only what exists can cause: An intrinsic view of free will', arXiv, (2206.02069), 1-24.

Tyler, Cowen. and Soutwood, Ben (2019), 'Is the rate of Scientific Progress Slowing Down?', GMU Working Paper in Economics, 21 (13), 1-46.

"Alexa, ti piace la Nutella?": cosa impariamo dal rapporto tra bambini e smart speaker

L'introduzione degli smart speaker in famiglia si accompagna a una redistribuzione delle capacità di azione di tutti gli attori coinvolti. Ma cosa succede alle relazioni genitore-figlio?

Di **Giovanna Mascheroni**, Università Cattolica del Sacro Cuore

Gli smart speakers non sono stati progettati e commercializzati per un pubblico di bambini – almeno fino al lancio del primo **Echo Dot Kids Edition** negli Stati Uniti nel 2018 e, a partire dal 2021, anche in Europa.

Eppure, gli smart speaker sono oggi presenti in molte case e, di conseguenza, nelle vite di molti bambini in età pre-scolare (Rideout & Robb, 2020; Wald et al., 2023; Wang, Luo & Wang, 2023). In Italia, un'indagine condotta a settembre 2020 su un campione rappresentativo di genitori di almeno un figlio di età pari o inferiore agli 8 anni (Mascheroni e Zaffaroni, 2021) ha rilevato la presenza di smart speakers nel 46% delle famiglie partecipanti: in queste famiglie, un bambino su tre interagisce con lo smart speaker in autonomia, senza interventi da parte dei genitori. Fra i genitori, invece, il 43% dichiara di usare lo smart speaker per raccontare le favole della buona notte ai figli.

Le ricerche sull'interazione fra bambini e smart speakers

Seppur ancora in una fase esplorativa, la ricerca sull'interazione fra bambini e smart speaker nel contesto domestico ha già evidenziato alcuni pattern ricorrenti. Innanzitutto, gli smart speakers sono di solito posizionati in uno spazio comune come il salotto e la cucina, e connotati come medium familiare (Lopatovska et al., 2019). Molti studi rilevano, inoltre, una diminuzione nell'uso rispetto alle prime fasi di addomesticamento, caratterizzate da un effetto novità. Si riscontrano anche alcune **differenze generazionali** nelle pratiche d'uso: se gli adulti usano Alexa o Google Home soprattutto per l'ascolto di musica, il controllo di dispositivi di domotica, e la ricerca di informazioni, oltre all'ascolto musicale i **bambini prediligono pratiche ludiche**, come la richiesta di barzellette o la conversazione (Lopatovska et al., 2019; Garg & Sengupta, 2020). Inoltre, i bambini di età inferiore ai sette anni hanno più probabilità di attribuire un'identità umana agli agenti conversazionali (Garg & Sengupta, 2020). Infatti, come teorizzato dagli studiosi di Human Machine Communication (HMC), gli agenti conversazioni sono programmati per entrare nella relazione comunicativa nel ruolo di partner comunicativi (Guzman, 2018, 2019; Guzman & Lewis, 2020)., appunto, i cui tratti umani si manifestano nel genere e nel ruolo sociale: quello di un assistente (Guzman, 2019) o di un servitore (Fortunati et al., 2022). La loro percezione, da parte degli utenti, oscilla fra il polo della macchina o, viceversa, dell'interlocutore quasi-umano, con i bambini più inclini ad aderire all'universo simbolico della simulazione della conversazione umana.

La letteratura ha anche messo a fuoco le possibilità di empowerment dei bambini in età prescolare offerte dall'interazione vocale, che permette ai più piccoli l'accesso diretto a contenuti mediali come canzoni e favole, senza la necessità di digitare su una tastiera (Beneteau et al., 2020).

L'introduzione degli smart speakers in famiglia si accompagna, quindi, a una redistribuzione delle capacità di azione di tutti gli attori coinvolti: lo smart speaker, innanzitutto, che acquisisce una certa autonomia nell'esecuzione di compiti, dal controllo delle interfacce domotiche connesse, alla selezione dei contenuti mediali (Mascheroni & Siibak, 2021); i genitori che, se da un lato vedono diminuire la propria capacità di controllo sull'accesso dei più piccoli ai contenuti mediali, dall'altro possono esercitare forme di genitorialità "aumentata" o assistita dallo smart speaker (Beneteau et al., 2020) – usato come mediatore neutrale nei conflitti o rinforzo degli ordini (Wang, Luo & Wang, 2023), ad esempio ricordando ai bambini che è ora di andare a letto, o di spegnere il tablet per andare a giocare al parco; infine i bambini che, come anticipato, possono aggirare le regole sull'uso degli schermi, chiedendo allo smart speaker di accendere la televisione sulla serie preferita.

Bambini e smart speakers in Italia

Come parte del progetto DataChildFutures, finanziato dalla Fondazione Cariplo nell'ambito del Bando Ricerca Sociale 2019, abbiamo condotto una ricerca qualitativa longitudinale nell'arco di 16 mesi che ha coinvolto 20 famiglie con almeno un figlio di età pari o inferiore a 8 anni. Il campione, costruito attraverso **un campionamento a scelta ragionata**, è vario sotto il profilo sociodemografico, della struttura famigliare (con 4 famiglie di genitori separati, 8 famiglie con figli unici, due famiglie numerose -rispettivamente con 7 e 4 figli), origine etnica, partecipazione religiosa. Fra le famiglie del nostro campione, 13 hanno o hanno avuto uno smart speaker, in particolare: 9 famiglie continuano a usare uno smart speaker; 2 famiglie hanno smesso di usare lo smart speaker per ragioni legate alla privacy, o, come vedremo, per la natura dell'interazione fra il figlio e l'agente conversazionale; infine, in 2 famiglie separate, i bambini hanno accesso a uno smart speaker solo a casa del padre.

Anche nel nostro campione, le pratiche d'uso rientrano nelle **categorie dell'intrattenimento** (musica, favole, barzellette), dell'informazione (notizie, previsioni del tempo, traduzioni dall'inglese o altre informazioni per i compiti scolastici, ecc.), dell'automazione (soprattutto per il controllo delle luci smart). Solitamente, tali pratiche non vanno a sostituire integralmente l'uso di altri media: la pratica dell'ascolto musicale, ad esempio, si lega a diversi dispositivi e piattaforme (smart speaker, tablet, radio, YouTube, Spotify ecc.) a seconda del contesto e del momento della giornata.

Una nuova autonomia per i bambini?

Come già rilevato nelle ricerche condotte negli Stati Uniti, l'interazione con gli assistenti vocali è un'occasione, anche per i bambini in età pre-scolare, di accedere direttamente a contenuti mediali, spesso disattendendo le regole dei genitori. L'accresciuta autonomia raggiunta si traduce, quindi, nel controllo, da parte del bambino, non solo dello smart speaker, ma anche di altri dispositivi connessi, ad esempio la smart TV. Se queste dinamiche riconfigurano le relazioni di potere fra genitori e figli, tuttavia non si traducono automaticamente nella **perdita di potere da parte del genitore**. Infatti, all'emancipazione del bambino spesso corrisponde un'emancipazione del genitore, che può dedicarsi al lavoro o a faccende domestiche mentre il figlio si intrattiene con le favole raccontate dall'agente conversazionale, come racconta la mamma di un bambino di 4 anni. In tal senso, **Alexa e Google incarnano la versione senza schermo delle babysitter digitali** come tablet, smartphone e televisione (Elias & Sulkin, 2019; Haddon & Holloway, 2018; Mascheroni e Zaffaroni, 2023; Nikken, 2022). Proprio l'assenza di schermo costituisce, anzi, una motivazione centrale nella scelta di comprare uno smart speaker. Ad esempio Letizia, mamma separata di un bambino di 6 anni, racconta di aver comprato Alexa per evitare che Ludovico dovesse usare il suo

cellulare ogni volta che aveva voglia di ascoltare musica: “L'avevo comprato con molta leggerezza perché pensavo appunto che potesse essere, anche per lui, una cosa comoda perché appunto non doveva accedere al mio cellulare, non doveva... E allora vuoi sentire la musica? La puoi sentire, vuoi sentire la tua musica? La puoi sentire, puoi chiedere, puoi richiedere, puoi fare una ricerca. "Voglio sapere qualcosa". Allora è più facile”.

In altri casi, invece, **l’empowerment guadagnato dal bambino genera conflitti per il controllo dello smart speaker,** dove la posta in gioco non è tanto l’affermazione di un’identità autonoma attraverso i propri gusti musicali, quanto la sfida ai genitori e alle norme valoriali del nucleo famigliare. Si tratta di pratiche di “accesso interrotto” (“disrupted access”, Beneteau et al., 2020), come racconta Gabriella, mamma di un bambino di 6 anni: “Io lo avvio ma poi ascoltano tutti, cioè, nel senso che comunque... Poi è un po’ una lotta un po’ con Guido, ogni tanto non vuole ascoltare della musica che noi ascoltiamo… ma anche se gli piace. Un po’, a volte, magari si emoziona troppo e preferisce non ascoltarla in quel momento quella canzone, poi magari la va a ricercare in un secondo momento […] E quindi dice: “Google, basta.” e chiude. Non vi dico le liti!”

Anche i conflitti e le negoziazioni che nascono fra genitori e figli in relazione all’accesso ai media digitali non sono una novità, anzi, hanno accompagnato l’“addomesticamento” di ogni innvazione tecnologica. Caratteristico degli smart speakers, tuttavia, è il fatto che le dinamiche di resistenza alle, e rinegoziazione delle relazioni di potere consolidate avviene sotto forma di interazione comunicativa con lo smart speaker, anziché nella forma più tradizionale di un’interazione comunicativa fra genitori e figli intorno a certi media.

La comunicazione con gli smart speakers

Il tratto distintivo degli smart speakers, e degli agenti conversazionali in generale, è quello di essere **forme di “comunicazione artificiale”** (Esposito, 2022), vale a dire “media (in parte) automatizzati e (in parte) autonomi che servono da interfacce di (quasi-)comunicazione con gli esseri umani” (Hepp, 2020, p. 1416). Anche se la relazione comunicativa con gli smart speakers tende a diminuire nel tempo – quando svanisce l’effetto novità – le famiglie si trovano a interagire con una voce connotata in termini di genere e programmata per eseguire autonomamente certi compiti, ma anche per rispondere a curiosità e conversare con l’utente.

L’interazione comunicativa con gli smart speakers viene modellata a partire dal processo di **attribuzione di senso** da parte del partner comunicativo umano. Più precisamente, cruciale a determinare la riuscita o il fallimento dell’atto comunicativo è l’attribuzione di una natura umana o, in alternativa, di una macchina allo smart speaker: da qui, infatti, risulta l’efficacia percepita della risposta comunicativa - vale a dire, quanto lo smart speaker si conforma con le aspettative dell’interlocutore. In linea con le ricerche internazionali, la tendenza ad antropomorfizzare lo smart speaker, e a testare la sua natura, è visibile soprattutto fra i bambini più piccoli. Infatti, i bambini che hanno partecipato alla nostra ricerca chiedono a Alexa se le piace la Nutella (Elisa, 8 anni) o se ha dei genitori (Alessandro, 5 anni). **L’attribuzione di caratteristiche antropomorfe e di una personalità proprio è fortemente influenzata dai tratti di genere con cui lo smart speaker è stato programmato**. Sia i bambini che i loro genitori riconoscono che Alexa e Google hanno una diversa identità di genere, anche se le interpretazioni di tali identità non sempre coincidono: infatti, se tendenzialmente Alexa è percepita come più simpatica ma anche più intelligente in quanto donna, non manca chi pensa che sia invece Google Home a superare Alexa in intelligenza perché “Google è un maschio, e sa più cose” (Carlotta, 7 anni).

L'antropomorfizzazione dello smart speaker, come anticipato, genera nell'utente aspettative di reciprocità che vengono spesso disattese. Di conseguenza, anche se l'antropomorfizzazione di Alexa o Google è fonte di preoccupazione per qualche genitore, in realtà i bambini finiscono spesso per considerare lo smart speaker "stupido" perché non capisce le loro richieste e non risponde come si aspetterebbero (come un essere umano).

La capacità di azione degli smart speakers

Gli smart speakers non sono solo partner comunicativi: sono pienamente inseriti nelle infrastrutture della datificazione che colonizzano il domestico e il privato (Couldry & Mejias, 2019; Mascheroni & Siibak, 2021). **Gli smart speakers introducono un nuovo livello di intermediazione nell'accesso dei contenuti che è fortemente "data-driven":** la selezione algoritmica dei contenuti più adatti ai noi avviene in maniera opaca, nella forma di una risposta vocale: una voce di cui ci fidiamo perché sa cosa ci piace e cosa cerchiamo. La selezione algoritmica dei contenuti altera le dinamiche di potere fra uomo e macchina: come riconoscono molti genitori, infatti, la capacità di scelta risulta ridotta a priori. Durante una delle nostre visite, Petra, mamma di un bambino di 4 anni e di uno di 18 mesi, incoraggia il figlio maggiore a chiedere una storia a Google, per mostrare ai ricercatori la sua autonomia di interazione con lo smart speaker. Quando Google seleziona Rapunzel dalla playlist di favole di Spotify, la mamma commenta che, dal momento che il bambino ha spesso chiesto Rapunzel, Google ora propone sempre la stessa favola. L'agency della macchina limita l'agency umana, riducendo una lista di contenuti potenzialmente infinita a un numero ristretto di opzioni selezionate in base alle pratiche d'uso passate, e alle pratiche di chi è classificato come simile.

Conclusioni

In conclusione, se le relazioni genitore-figlio ricalcano modalità già osservate nei processi di addomesticamento dei media digitali (il ricorso, da parte dei genitori a babysitter digitali, o conflitti iniziati dai figli per reclamare autonomia nell'accesso ai media), la capacità di azione degli smart speaker ridefinisce l'agency umana. Questa ridefinizione avviene grazie alla natura degli smart speakers come macchine dalla voce umana: detto altrimenti, il posizionamento nella relazione comunicativa come veri e propri partner comunicativi permette di nascondere, e normalizzare, la sottrazione di agency.

Bibliografia

Beneteau, E., Boone, A., Wu, Y., Kientz, J., Yip, J., & Hiniker, A. (2020). Parenting with Alexa: Exploring the introduction of smart speakers on family dynamics. CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, April 21, 1–13. https://doi.org/10.1145/3313831.3376344

Elias, N., & Sulkin, I. (2019). Screen-assisted parenting: The relationship between toddlers' screen time and parents' use of media as a parenting tool. Journal of Family Issues, 40(18), 2801-2822. https://doi.org/10.1177/0192513X19864983

Esposito, E. (2022). Comunicazione artificiale: Come gli algoritmi producono intelligenza sociale. Egea.

Fortunati, L., Edwards, A., Edwards, C., Manganelli, A. M., & de Luca, F. (2022). Is Alexa female, male, or neutral? A cross-national and cross-gender comparison of perceptions of Alexa's gender

and status as a communicator. Computers in Human Behavior, 137, 107426. https://doi.org/10.1016/j.chb.2022.107426

Garg, R., & Sengupta, S. (2020). He is just like me: A study of the long-term use of smart speakers by parents and children. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4(1), 1–24. https://doi.org/10.1145/3381002

Guzman, A. L. (2018). What is human–machine communication, anyway? In L. Guzman (Ed.), Human–machine communication: Rethinking communication, technology, and ourselves (pp. 1–28). Peter Lang.

Guzman, A. L. (2019). Voices in and of the machine: Source orientation toward mobile virtual assistants. Computers in Human Behavior, 90, 343–350. https://doi.org/10.1016/j.chb.2008.03.008

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A Human-Machine Communication research agenda. New Media & Society, 22(1), 70-86. https://doi.org/10.1177/1461444819858691

Hepp, A. (2020). Artificial companions, social bots and work bots: communicative robots as research objects of media and communication studies. Media, Culture & Society, 42(7-8), 1410-1426. https://10.1177/0163443720916412

Lopatovska, I., Rink, K., Knight, I., et al. (2019). Talk to me: Exploring user interactions with the Amazon Alexa. Journal of Librarianship and Information Science, 51(4), 984–997. https://doi.org/10.1177/0961000618759414

Mascheroni, G., & Siibak, A. (2021). Datafied childhoods: Data practices and imaginaries in children's lives. Peter Lang.

Mascheroni, G., e Zaffaroni, L.G. (2023. From "screen time" to screen times: Measuring the temporality of media use in the messy reality of family life. Communications. https://doi.org/10.1515/commun-2022-0097

Mascheroni, G., e Zaffaroni, L.G. (2021). Bambini e intelligenza artificiale, come bilanciare i rischi: gli studi. Agenda Digitale. https://www.agendadigitale.eu/cultura-digitale/bambini-e-intelligenza-artificiale-tutti-i-diritti-in-gioco-e-come-bilanciarli/

Nikken, P. (2022). The touch-screen generation: Trends in Dutch parents' perceptions of young children's media use from 2012–2018. Communications, 47(2), 286-306. https://doi.org/10.1515/commun-2020-0028

Rideout, V., & Robb, M. B. (2020). The Common Sense census: Media use by kids age zero to eight, 2020. Common Sense Media. www.commonsensemedia.org/research/the-common-sense-census-media-use-by-kids-age-zero-to-eight-2020

Wald, R., Piotrowski, J. T., Araujo, T., & van Oosten, J. M. (2023). Virtual assistants in the family home. Understanding parents' motivations to use virtual assistants with their Child (dren). Computers in Human Behavior, 139, 107526. https://doi.org/10.1016/j.chb.2022.107526 ù

Wang, B., Luo, L., & Wang, X. (2023). “Back to the living room era”: Smart speaker usage and family democracy from the family dynamic perspective. New Media & Society. https://doi.org/10.1177/14614448231155624

Ripensare il rapporto tra intelligenza umana e artificiale, per una "ecologia gestaltica" dell'AI

Discutere se sia possibile ipotizzare che accanto a una soggettività umana ci sia una soggettività "AI" è poco utile. Conviene piuttosto considerare il contributo dell'intelligenza artificiale a una visione relazionale della realtà ed eventualmente provare ad approfondirne il carattere intrinsecamente relazionale

Di Salvatore Tedesco, Università di Palermo

«Il mondo dei contenuti concreti ha carattere di gestalt» scriveva **Arne Næss** promuovendo il progetto di una ecologia profonda, in grado di ripensare **il ruolo del soggetto umano nel quadro di una realtà concepita come interazione molteplice fra differenti soggetti biologici e costituenti dell'ambiente naturale, artificiale, simbolico**.

Cosa accade nel momento in cui una simile ecologia gestaltica viene ripensata alla luce di **un quadro più inclusivo**, in cui le interazioni ambientali (il fare e il ricevere, gli assetti senso-motori e le operazioni di comprensione, ristrutturazione, progettazione della realtà) riguardano insieme agenti naturali e dispositivi di AI? Soprattutto: cosa accade nel momento in cui in tale quadro operativo si fa spazio per interazioni capaci di implicare una "creatività" non solo nell'adozione, ma nell'ideazione delle stesse "regole"?

https://www.agendadigitale.eu/cultura-digitale/linformatica-ha-rotto-il-potere-dellintelligenza-umana-ecco-perche-parliamo-di-rivoluzione/

Così rappresentato, il "mondo dei contenuti concreti" è evidentemente **il mondo della nostra esperienza concreta quotidiana**, il mondo delle quotidiane **interazioni** in cui si sviluppa e si struttura la nostra vita. Il modello concettuale di cui parliamo ci suggerisce dunque di **intendere la realtà del nostro mondo come un ambiente relazionale**, intendendo con questa espressione un ambiente in cui non esistono "cose separate" – per esempio soggetti umani contrapposti ad oggetti naturali e artificiali – ma relazioni gestaltiche, cioè **relazioni configurate**, eventi dotati di una forma, di un modo di presentarsi e di una composizione interna. In ragione di ciò potremmo dire che gli elementi della realtà possono appunto essere compresi solo a partire dai tipi di relazione che si costruiscono fra di loro.

Un modello teorico di questo tipo, concepito negli ultimi decenni come approfondimento di un atteggiamento ecologico, esprime indubbiamente una forte critica nei confronti dello sfruttamento della natura da parte dell'essere umano, ed apre invece a una considerazione basata appunto sulla **cooperazione**, l'interazione fra "agenti" umani, naturali, artificiali all'interno di uno spazio (chimico-fisico, ma anche sociale, culturale, simbolico) in continua trasformazione. Parliamo

dunque di **sistemi viventi ed elementi che si pongono evidentemente come differenti**, ma coinvolti nella stessa opera di costruzione della realtà ed eventualmente minacciati dagli stessi squilibri.

Risulta di estremo interesse a parere di chi scrive il fatto che un modello teorico di questo tipo – dunque il modello di una ecologia profonda, che non si limita a lanciare qualche appello di natura etica, ma si pone come un piano di descrizione della realtà adeguato alla sua effettiva costituzione e trasformazione – proprio per la sua natura relazionale, possa essere utilmente adottato per descrivere un quadro "esteso" quale sempre più risulta essere quello della nostra esperienza, in cui **l'interazione che caratterizza il nostro ambiente si caratterizza per la presenza non solo di componenti artificiali**, di dispositivi tecnici/tecnologici di vario tipo e di relazioni simboliche (come è noto tutto questo fa parte in ultima analisi di ogni cultura umana, anche delle più arcaiche fra quelle attestate), ma in senso specifico per il ricorso all'intelligenza artificiale, che appunto appare interagire in modo sempre più pervasivo con i più diversi aspetti del nostro mondo quotidiano, delle nostre scelte e preferenze, persino della costruzione di valori o disvalori condivisi ecc.

Gli ambiti in cui si realizza l'interazione "gestaltica"

Forse insomma risulta in ultima analisi poco utile discutere se sia possibile ipotizzare che accanto a una soggettività umana ci sia **una soggettività "AI"**, per il semplice motivo che quell'idea di un "soggetto sovrano" contrapposto a degli oggetti è in se stessa poco utile descrittivamente, e addirittura rappresenta una scelta teorica distruttiva dal punto di vista delle sue implicazioni operative.

Conviene piuttosto considerare il contributo dell'intelligenza artificiale a **una visione relazionale della realtà**, ed eventualmente provare ad approfondire il carattere intrinsecamente relazionale dell'intelligenza artificiale.

Quanto al primo punto, proviamo brevemente e senza alcuna pretesa di completezza a indicare gli ambiti in cui si realizza l'interazione "gestaltica" di cui si diceva:

Qualità vs quantità

Parliamo di una realtà di relazioni che non andranno descritte in primo luogo in termini quantitativi "oggettivi", e non perché si tratti di una esperienza solo "soggettiva", ma perché si tratta appunto di correlazioni che creano un ambiente insieme con le "regole" di funzionamento dei diversi agenti ed elementi che ne fanno parte. Torneremo nell'ultima parte delle nostre brevi considerazioni a questa correlazione fra ambiente e regole. Osserviamo intanto che in questo senso:

La forma è sempre relazionale

Le forme sono "costellazioni irriducibili" i cui elementi appunto hanno modalità di comportamento, modalità di esistenza, specificamente descritte dalle forme cui danno luogo; per dirla con Olaf Breidbach, considerando che ogni forma/Gestalt necessariamente si trasforma, diviene via via differente, «una specifica Gestalt determina anche cosa rimane invariante al di sotto delle trasformazioni […]. Dal momento che ogni pattern si trasforma come un intero, tali invarianti a loro volta consistono di relazioni interne alla stessa Gestalt». Per questo potremmo aggiungere che:

Le relazioni gestaltiche si esplicitano in quanto rilevanza per la costruzione, il mantenimento e la trasformazione dello stesso ambiente.

Si tratta dell'aspetto decisivo per intendere il carattere "qualitativo" del discorso. Le interazioni che hanno luogo in un ambiente, in una realtà concretamente sperimentata, muovono per esempio dalla **rilevanza biologica della relazione che si viene a creare**, e non hanno carattere quantitativo, ma qualitativo: la relazione muove da una aspettativa carica di promesse positive o negative, minacce o attese, impulso all'apertura relazionale o bisogno di chiusura nei confronti della stessa relazione, ecc. La costruzione di una rete di relazioni, il suo mantenimento e le sue trasformazioni sono legate appunto a queste "qualità" colte nella relazione e infine nell'ambiente stesso. In questo senso le relazioni gestaltiche si esplicitano in quanto modalità di percezione, forme di movimento, schemi di comprensione, progettazione e ristrutturazione dell'ambiente.

Altrettanto importante è il fatto che **le relazioni gestaltiche possano esplicitarsi per così dire in relazione a "se stesse"**, che abbiano cioè un livello riflessivo, che si traduce in relazioni "sociali", culturali, simboliche; ciò implica un discorso metodologico che è al tempo stesso interno al singolo ambito disciplinare e però anche capace di apertura transdisciplinare (cioè filosofica).

L'intelligenza artificiale come agente che interviene nella costruzione dell'ambiente

In riferimento ai punti qui descritti, non si può fare a meno di rilevare che l'intelligenza artificiale appare estendere la configurazione gestaltica propria dell'ambiente nel suo carattere relazionale proprio perché **a differenza di altre tecnologie l'intelligenza artificiale non è meramente uno "strumento" a disposizione di un soggetto che in questo modo possa potenziare il proprio controllo sulla realtà**; l'intelligenza artificiale si propone piuttosto come un ulteriore agente che interviene nella costruzione dell'ambiente e dunque della realtà nelle sue varie articolazioni.

Si tratta di un punto evidentemente decisivo per la comprensione del ruolo dell'intelligenza artificiale, e non a caso questa è appunto la componente che da sempre suscita le maggiori attese e attorno alla quale si raccolgono anche i maggiori timori o le fantasie distopiche, da Hal 9000 di Odissea nello spazio sino ai **dibattiti** di questi mesi su ChatGPT e Dall-E.

Laddove però – sottolineo ancora – l'idea che sia possibile attribuire all'intelligenza artificiale una "soggettività" classica, e dunque la presenza di un "Io" in senso forte, rischia di basarsi in ultima analisi su un paradigma filosoficamente assai compromesso e di costruire su quella base delle attese e delle esigenze assolutamente problematiche, puntare l'accento sul carattere relazionale/ambientale tanto dei "soggetti biologici" (umani e non) quanto dei "soggetti AI" (come forse potremmo provare a definire secondo questa accezione relazionale i dispositivi/programmi basati sull'intelligenza artificiale), significa probabilmente coglierne in modo più adeguato tanto il funzionamento nella nostra realtà, quanto appunto la caratterizzazione intrinseca, cui si potrà riferire il contributo – che definirei davvero innovativo – alla **descrizione e alla comprensione di quello che c'è nel nostro mondo**, del suo piano ontologico per dirla in termini che rinviano alla proposta formulata da Arne Næss di una "Gestalt Ontology" che si accompagna a una ecologia gestaltica.

Memoria e attenzione relazionale dell'AI

Mi limito a saggiare brevemente due aspetti della questione, forse però decisivi: anzitutto, è senz'altro vero che **appare problematico attribuire all'intelligenza artificiale una intenzionalità**; l'esclusione di questa caratteristica è spesso apparsa confermare il ruolo puramente

strumentale dell'intelligenza artificiale, alla quale verrebbe dunque attribuito un ruolo di servizio nei confronti di una soggettività classica (ed esclusivamente umana, ovviamente), dotata di volontà autoconsapevole, di intenzionalità, e specularmente contrapposta a oggetti sui quali esercitare il proprio giudizio e la propria legislazione.

Il punto di vista gestaltico e ambientale prima espresso cambia però non poco le cose, relativizza il ruolo dell'intenzionalità, e per esempio ci può portare a chiederci se piuttosto l'intelligenza artificiale non sia caratterizzata da quella che potremmo provare a definire in senso descrittivo e non psicologico come memoria ed attenzione relazionale, prestazione a partire dalla quale la possibilità di processare una quantità enorme di dati mettendoli in circolo nell'ambiente relazionale non vale tanto come "imitazione" di una facoltà tipica dell'essere umano o comunque di un organismo vivente, ma **costituisce in senso proprio un contributo peculiare alla costruzione, al mantenimento e alla trasformazione dell'ambiente relazionale**. Si tratterebbe, secondo il punto di vista che qui si propone, di un contributo in ultima analisi irriducibile, non riferibile all'attività di altri elementi dello stesso ambiente relazionale. L'essere umano, in questo senso, si pone a tutti gli effetti in dialogo con l'intelligenza artificiale all'interno del contesto ambientale relazionale, all'interno dell'irriducibilità della specifica Gestalt.

Le GAN e il carattere relazionale dell'ambiente gestaltico

Non meno rilevante risulta in questa prospettiva il secondo elemento che ci proponiamo di saggiare: si è già detto che una delle caratteristiche più importanti sviluppate dalle relazioni gestaltiche che stiamo considerando è la possibilità che la configurazione/Gestalt sia non solo "messa in forma", sperimentata, ma che in essa si sviluppi una riflessività, per la quale la forma acquisti una dimensione simbolica e la relazionalità stessa si sviluppi come discorso di metodo che verte per così dire sul modo di stare nella relazione. È probabilmente per questo che la filosofia e l'estetica (penso agli ottimi studi, in Italia, di Alice Barale) dedicano oggi particolare attenzione alle **reti generative avversarie (GAN)** su cui si basano molte fra le attuali configurazioni emergenti dell'intelligenza artificiale. Si tratta notoriamente di reti composte da un generatore e un discriminatore che operano letteralmente "l'uno contro l'altro" in ragione di una architettura di base che Michael Castelle definisce come «una caratteristica struttura interattiva e duale». Ciò significa appunto che la generazione di immagini visive, di forme sonore o di discorsi verbali cui esse danno luogo è frutto della loro intrinseca relazionalità.

Potremmo anche dire pertanto che **il carattere relazionale dell'ambiente gestaltico viene "introiettato" dalle GAN**, nelle quali dunque – senza bisogno di perdersi a ipotizzare una "consapevolezza umana" del tutto fuori luogo – si sviluppa in senso proprio una dimensione di riflessività che rinvia al carattere relazionale dell'ambiente gestaltico, una sorta di relazionalità al quadrato, se vogliamo, che sta alla base del modo di funzionare proprio delle GAN, e ne costituisce il contributo alla relazione complessiva all'interno dell'ambiente gestaltico.

L'interazione uomo/macchina alla base di un ripensamento complessivo del nostro ambiente

Per essere ancora più chiari: l'interazione uomo/macchina che si realizza ad esempio negli interventi "artistici" o artistico/ludici possibili con Dall-E, anziché esser vista come un "non ancora" che dia luogo a commenti più o meno improntati al pathos dell'unicità umana e/o dell'ormai prossimo infrangersi di questo mito ("la macchina non ha ancora imparato a pensare da sola"; "la macchina non esautora mai il ruolo insostituibile dell'essere umano"; "la macchina non ha esperienza corporea"; ecc.), appare forse assai più prosaicamente – ma forse anche in modo tale da

marcare davvero un inedito nella nostra realtà – come una ristrutturazione e un ripensamento complessivo del nostro ambiente e delle forme che in esso possono aver luogo: dell'ecologia e dell'ontologia gestaltica al tempo stesso, appunto.

In questo senso il tema sarebbe quello del fare uso (penso alla riflessione sul concetto di uso da Wittgenstein a Paolo Virno) di un ambiente gestaltico. Tema che qui posso solo sfiorare, ma che evidentemente si declina a partire dalla prossimità fra gli elementi ambientali, dall'individuare non tanto le proprietà essenziali di un determinato soggetto/oggetto, ma piuttosto una appropriatezza nelle modalità di approccio relazionale che sarà evidentemente sempre in divenire e andrà considerata su una molteplicità di piani di discorso, da quello funzionale, a quello delle risorse, sino al piano delle scelte etiche. Così considerate, le regole di funzionamento, le regole relazionali, sarebbero da intendere meno come un set di istruzioni determinate che non come **un "insieme mobile" di possibilità di interazione.**

Questa considerazione ci guida verso l'ultimo passo della nostra breve ricognizione: possiamo descrivere il fare uso di un ambiente gestaltico come l'eseguire in esso dei gesti capaci di attivare le procedure relazionali più appropriate alla situazione. Questa semplice considerazione ci guida direttamente verso uno dei temi più frequentati nel dibattito sull'intelligenza artificiale, specialmente in relazione all'arte, ma non solo: in ogni caso il discorso estetico/artistico può servire qui da momento di anticipazione/sperimentazione di un problema più diversificato.

Ai e creatività

Mi riferisco alla questione della creatività in riferimento all'intelligenza artificiale, dalle analisi di Margaret Boden in avanti. Ed anzitutto: esiste? È in generale possibile parlarne, oppure ci troviamo in ultima analisi di fronte una **ricombinazione dei dati forniti al sistema,** eventualmente per noi imprevista solo perché siamo empiricamente incapaci di processare mentalmente la stessa mole di dati della macchina?

Anche qui forse l'assetto ambientale/relazionale proposto può fornire un punto di vista almeno parzialmente differente, appunto in relazione ai concetti pocanzi introdotti di uso e di gesto.

Pensiamo dunque al gesto come "attivazione" di determinate procedure relazionali appropriate alla situazione determinata: a grandissime linee potremmo configurare tre modalità differenti.

Potremmo cioè immaginare **un set di istruzioni "rigide"** che il determinato gesto meramente applica alla situazione in oggetto: quando premo col mouse il pulsante 'corsivo' del mio programma di videoscrittura, il programma applica una regola precisa consentendomi appunto di scrivere applica in corsivo. Si tratta evidentemente di una risposta non dotata di creatività.

Potremmo poi immaginare una creatività a bassa intensità, per così dire, nel momento in cui invece il gesto non si limita ad applicare una regola fissa, ma piuttosto – come ci siamo espressi più su – attiva una procedura, andando alla ricerca della risposta più adatta alla specificità della situazione relazionale.

Credo che ci siano ormai pochi dubbi sul fatto che le reti basate sull'elaborazione antagonista (GAN) abbiano in questo senso accesso a un tale tipo di "creatività".

Conclusioni

È però possibile immaginare che il gesto non abbia solo la funzione di attivare un certo comportamento all'interno dell'ambiente gestaltico (certo contribuendo alla ridefinizione della stessa Gestalt, alla sua progressiva modificazione/adattamento), ma che per così dire rappresenti rispetto a quell'ambiente uno scarto, un modo di "andare altrove", che implica l'accesso a nuove e differenti modalità procedurali, che implichi una trasformazione "drammatica" della Gestalt relazionale. Potremmo parlare in questo caso di creatività ad alta intensità, creatività in cui il gesto investe appunto la Gestalt in quanto tale.

Credo, conclusivamente, che questa soglia ulteriore rimanga una vera frontiera di estremo interesse, non certo per pensare un modo di esautorare l'umano, ma per ripensare il modo in cui l'intelligenza umana e l'intelligenza artificiale cooperano e si relazionano nella realtà, per ripensare gli orizzonti che via via vi si configurano, e chiaramente le scelte che siamo e saremo chiamati ad operarvi.

Bibliografia

A. Barale (a cura di), Arte e intelligenza artificiale, Jaca Book, Milano 2020

M.A. Boden, La mente creativa, Mondadori, Milano 1995

O. Breidbach, J. Jost, On the gestalt concept, in "Theory in Biosciences", 125, 2006

M. Castelle, La vita sociale delle reti antagoniste generative (GAN), in A. Barale (a cura di), Arte e intelligenza artificiale, Jaca Book, Milano 2020

E. Garroni, Creatività, Quodlibet, Macerata 2009

M. Mazzeo, Il bambino e l'operaio. Wittgenstein filosofo dell'uso, Quodlibet, Macerata 2016

A. Næss, Ecology, community and lifestyle. Outline of an Ecosophy, Cambridge U.P., Cambridge 1989

A. Næss, The Selected Works of Arne Næss, a cura di H. Glasser, A. Drengson, 10 voll., Springer, Dordrecht 2005

A. Næss, Reflections on Gestalt Ontology, in "The Trumpeter", 21, 1, 2005

A. Næss, Siamo l'aria che respiriamo. Saggi di ecologia profonda, Piano B 2021

J. von Uexküll, Ambienti animali e ambienti umani, Quodlibet, Macerata 2010

P. Virno, L'idea di mondo, Quodlibet, Macerata 2015

I quaderni di

Agenda Digitale.eu

NETWORK DIGITAL 360

Network Digital360 è il più grande network in Italia di testate e portali B2b dedicati ai temi della Trasformazione Digitale e dell'Innovazione Imprenditoriale, con oltre 50 fra portali, canali e newsletter.
Ha la missione di diffondere la cultura digitale e imprenditoriale nelle imprese e pubbliche amministrazioni italiane e di fornire a tutti i decisori che devono valutare investimenti tecnologici informazioni aggiornate e approfondite.
Il Network è parte integrante di Digital360HUB, il polo di Demand Generation di Digital360, che mette a disposizione delle tech company un'ampia gamma di servizi di comunicazione, storytelling, pr, content marketing, marketing automation, inbound marketing, lead generation, eventi e webinar.

VIA COPERNICO, 38
20125 - MILANO
TEL. 02 92852785
MAIL: MARKETING@DIGITAL4.BIZ