Capire l'Intelligenza Artificiale per non esserne fagocitati

Franco Padella, Mario Carmelo Cirillo

Dal recruiting alla differenziazione dei prezzi di beni e servizi, già oggi viene utilizzata l'Intelligenza Artificiale. E nel futuro quanto saranno autonome le macchine nei passaggi decisionali? E quali problemi si porranno? Compendio ragionato del convegno "Intelligenze al confine tra biologico e artificiale".

In modalità per noi più o meno consapevole, i modelli di Intelligenza Artificiale (IA) stanno occupando vieppiù ampi settori della nostra vita. Partendo dagli assistenti vocali come Siri, i sistemi "intelligenti" si sono espansi ovunque ed è in atto un percorso che li vedrà entrare nei più disparati gangli sociali. La recente diffusione pubblica di Chat-GPT ha reso evidenti a tutti le grandi possibilità raggiunte dalla Intelligenza Artificiale, ora arrivata ad essere addirittura capace di una interazione uomo-macchina mediante l'utilizzo del linguaggio naturale.

In realtà l'IA ha una lunga storia alle spalle, fatta di alti e bassi, che parte negli anni '50-'60 del secolo scorso. Originariamente gli approcci allo sviluppo della IA erano di tipo simbolico, cioé basati su una rappresentazione esplicita della conoscenza, con l'implementazione di regole e sequenze logiche ben definite dalla programmazione umana. Caratteristica propria di tali approcci è la piena trasparenza degli algoritmi utilizzati. L'approccio simbolico ha prodotto risultati considerevoli in diversi campi, ed ha permesso lo sviluppo dei primi sistemi esperti. Nonostante ciò, la complessità nella gestione delle conoscenze e la difficoltà nel rappresentarne il contesto in forma digitale ha portato negli anni '70 ad una successiva stagnazione della metodologia, con un declino della ricerca sulla IA: un inverno della IA interrotto alla fine degli anni '90. A partire dagli anni 2000 le difficoltà riscontrate fino a poco prima vengono superate, "semplicemente" eliminando l'approccio simbolico di tipo logico e sostituendolo con i dati, l'enorme quantità di dati che si andava via via raccogliendo su Internet. All'approccio simbolico subentra così il cosiddetto apprendimento automatico, il quale utilizza una rappresentazione dei dati basata sul calcolo delle probabilità, partendo da quanto appreso in maniera automatizzata dalla globalità dei dati a disposizione. Pur essendo i modelli di funzionamento sottostanti molto complessi, l'elaborazione dei dati è determinata probabilisticamente, e quindi non è richiesta alcuna conoscenza reale dei contenuti. In sintesi: mentre gli approcci simbolici si basavano sulla rappresentazione esplicita della conoscenza e sul ragionamento basato su regole, gli approcci di *deep learning* attuali si concentrano sull'apprendimento dai dati, dei quali, utilizzando modelli probabilistici complessi, i nuovi sistemi sono in grado di riconoscere le regolarità statistiche.

Il cambio delle modalità di funzionamento ha prodotto risultati sorprendenti, con sistemi di IA che mostrano capacità di dare soluzione a problemi specifici estremamente rilevanti. Il prezzo correlato a tutto questo è l'opacità delle modalità attraverso le quali viene prodotta una decisione, cosa che rende molto difficile produrre un giudizio di valore, positivo o meno, sulla decisione presa. Sistemi di IA sono utilizzati in molti campi. Pensiamo ad esempio alla biologia, alla chimica computazionale, alla medicina, ai materiali, alle tecnologie di sostenibilità energetica ed ambientale e così via. In negativo, banalmente sul piano minuto delle transazioni commerciali, esistono sistemi di IA che permettono la differenziazione delle tariffe di costo dei biglietti aerei o di corse Uber in funzione delle condizioni puntuali dell'utente mappato al telefono (ad esempio sovraccosti per chi viene profilato come incalzato da impegni urgenti o col dispositivo che sta per scaricarsi). Su un piano di maggiore impatto negativo, non si può fare a meno di pensare ai droni utilizzati quali sistemi d'arma, campo in cui la tecnologia è forse presente da più tempo.

Qui appare importante sottolineare come la diffusa presenza della IA sta portando e porterà ancora a grandi cambiamenti nel mondo del lavoro, così come nella intera organizzazione sociale. L'automatizzazione di molte attività, gli aumenti di efficienza nelle produzioni, la ristrutturazione di interi settori economici, commerciali e sociali (si pensi a neanche troppo futuristici robot utilizzati come infermieri, colf o badanti) creerà criticità importanti in termini di perdita di posti di lavoro, cosa che sarà tutt'altro che semplice compensare in altri settori. Sul piano dei sistemi di governo, vengono sempre più spesso delegate ai modelli di IA scelte di tipo eminentemente politico, cosa che era impensabile fino a poco tempo fa. Si va dall'analisi demografica, economica

e sociale al fine di definizione delle politiche, all'automazione dei processi decisionali, quali ad esempio l'allocazione di risorse pubbliche o l'erogazione di servizi. Esistono sistemi per la predizione e la prevenzione, focalizzati sulle situazioni di emergenza ma anche sistemi ottimizzati sulla profilazione dei comportamenti di singoli cittadini e gruppi, come ad esempio i sistemi di sicurezza e sorveglianza (si pensi alla profilatura "in automatico" della propensione di soggetti a compiere reati).

A fronte di tutte le problematiche che la tecnologia fa emergere, si rafforza la sua umanizzazione, con la conseguenza che una buona parte di coloro che la utilizzano o ne sono coinvolti considerano realmente intelligenti i sistemi di IA. Questo rischia di disarmare completamente la capacità critica dell'interlocutore umano, che messo di fronte ad una macchina intelligente, e forse anche consapevole, facilmente le può delegare compiti di estrema importanza e delicatezza, in quanto portatrice di una capacità superiore, vera o presunta, con modalità che via via pongono la macchina sempre più al di fuori del controllo umano.

È necessario riportare la IA sotto un controllo e una continua verifica dell'umano, trattandosi di un oggetto che macchina è e tale rimane, al di là di ogni sua capacità di linguaggio corretto e suadente. Questo richiede anche e soprattutto una demistificazione di partenza, tale da farne emergere le opportunità (che comunque ci sono) insieme ai rischi (che altrettanto ci sono). È richiesto, seppure per grandi linee, uno sforzo di comprensione anche riguardo le modalità tecnologiche di funzionamento della IA, in modo che si renda possibile una interlocuzione e una discussione informata tra tutti gli attori coinvolti, siano essi scienziati o tecnici, sindacati, partiti, o organizzazioni sociali, o semplici cittadini. Ed è richiesto uno sforzo di educazione, che porti al superamento delle condizioni di diffuso analfabetismo tecnologico (primario e di ritorno) in cui ci troviamo. Oggi esiste una enorme sfasatura tra quanto il mondo della conoscenza tecnico-scientifica sviluppa e propone, quanto il mercato immette nel sociale e le modalità politico-filosofiche con cui vengono analizzate e affrontate le questioni legate allo sviluppo della IA ed alla rapidità dei cambiamenti che questa sta imponendo alla società.
Sono mondi che poco si parlano, e spesso quando si parlano appaiono destinati a non capirsi.

In questo senso meritoriamente il Gruppo Interdisciplinare su Scienza, Tecnologia e Società (GI-STS) dell'Area della Ricerca di Pisa del CNR ha organizzato lo scorso 20 settembre l'incontro seminariale "Intelligenze al confine tra biologico ed artificiale", dove l'implementazione della tecnologia è stata confrontata con i modelli dell'intelligenza biologica, con annessa la questione della possibilità che una IA possa far emergere qualcosa che somgli ad una coscienza di sé.

Il convegno, coordinato da Valentina Tozzini del CNR, si cala negli ambiti sopra descritti con due differenti relazioni espositive, una di carattere maggiormente tecnico-scientifico e l'altra più focalizzata sugli aspetti filosofici e sociali della IA. Le due relazioni, la prima di Giorgio Ascoli, neuroscienziato e professore alla *George Mason University* negli USA, la seconda di Daniela Tafani, filosofa e docente di Etica e politica dell'intelligenza artificiale all'università di Pisa, mettono a confronto le differenti modalità di approccio alla IA, punti di vista che, malgrado le difficoltà che pure emergono, dovranno necessariamente trovare il modo di convergere verso una visione condivisa.

Intelligenza artificiale e cervelli naturali: analogie, differenze e perché è importante capirle

L'intervento di Giorgio Ascoli apre sulla domanda: "I sistemi di intelligenza artificiale che sono stati sviluppati mostrano una capacità molto elevata[1], nel risolvere problemi. Esistono analogie con i cervelli biologici? I mammiferi, e comunque gli esseri umani, oltre all'intelligenza hanno capacità senziente[2]. E' possibile che le intelligenze artificiali sviluppino anche esse una capacità senziente, qualcosa che somigli ad un barlume di coscienza di sé?"[3].

Proviamo a descrivere analogie e differenze nel comportamento tra i sistemi di tipo biologico e gli algoritmi delle reti artificiali, che sono alla base dei *Large Language Models* (LLM) che hanno portato a Chat-GPT. Esemplificando al massimo, partiamo da una immagine proveniente dalla ripresa di una videocamera, che vogliamo far riconoscere alla IA con l'obiettivo di individuare un eventuale ladro che si muove carponi differenziandolo da un cane di passaggio[4]. Sovrapponendo una griglia all'immagine, questa è suddivisa in piccoli elementi, ognuno dei quali possiede proprie

definite caratteristiche, tipicamente forma, colore, posizione ecc. Tali singoli elementi vengono immessi in una rete neurale informatica[5]. Qui le informazioni vengono filtrate per passaggi successivi tra i nodi della rete, processo che permette la valutazione dell'importanza dell'informazione connessa all'elemento passante. Al termine dei processi di valutazione, la rete, addestrata su grandi basi di dati, scarta le informazioni ritenute non rilevanti e riassembla in forma propria quelle giudicate rilevanti, fino ad arrivare al giudizio finale, in questo caso relativo alla presenza o meno del ladro.

Nei modelli che operano sul linguaggio[6] gli elementi costitutivi assemblano più insiemi di reti neurali, ognuno dei quali connesso ad uno scopo specifico. Esemplificando sulla traduzione di un testo, avremo un insieme di reti neurali dedicato a far comprendere il testo in ingresso[7] (codificatore) e un insieme di reti neurali dedicato alla produzione del testo in uscita[8] (decodificatore). Poiché il testo in ingresso come quello in uscita sono sequenze logiche di parole, sia la codifica che la decodifica utilizzano meccanismi di valutazione dell'importanza delle singole parole così come della probabilità di connessione di ognuna di queste con le altre presenti nel testo in esame (autoattenzione)[9]. Anche qui, è l'addestramento delle reti su larghissime basi di dati che permette di ottenere il risultato corretto[10] .

Differentemente da una rete neurale naturale[11], la struttura delle reti artificiali non cambia né nel tempo né nello spazio, ed il passaggio dell'informazione è sempre e solo unidirezionale verso gli strati adiacenti successivi (*feed forward*). L'apprendimento dei sistemi deriva da istruzioni statiche, avviene in tempi successivi al passaggio dell'informazione e al seguito di una attività di supervisione esterna alla rete.

Al di là di molteplici altri importanti dettagli costitutivi, già tutto questo differenzia in maniera sostanziale una rete di origine artificiale da una di tipo naturale. La rete neurale biologica può cambiare la propria configurazione nello spazio e nel tempo, l'informazione ha passaggi non necessariamente prefissati e che possono avvenire anche tra strati lontani. L'apprendimento è contemporaneo al passaggio dell'informazione e non necessita di una supervisione esterna.

Sono importanti le differenze per l'emergere di una coscienza? Secondo la Teoria dell'Informazione Integrata[12], in corrispondenza di una determinata percezione (sensoriale, di quel che accade, di sé) ci deve essere un meccanismo nell'architettura di rete che le corrisponda. Questo consiste in una relazione di causa-effetto auto-generata all'interno della rete, e solamente una relazione auto-generata è in grado di produrre informazione integrata non nulla (e quindi percezione). Mentre una rete neurale biologica è ben in grado di generare informazione al suo interno, tutti i risultati che derivino osservazioni che avvengono all'esterno della rete neurale[13] producono solo e sempre una quantità di informazione integrata nulla. Le configurazioni di rete monodirezionali che sono alla base dei LLM non sono quindi in grado di fare emergere una coscienza. Per quanto prodigiose, tali macchine sono e rimangono dei pappagalli stocastici[14].

A complicare le cose arriva ora la domanda: se accadesse, potremmo riconoscere una eventuale emersione di coscienza in una macchina? Questo risulta più difficile, in quanto è possibile simulare attraverso reti monodirezionali qualunque altra rete di informazione, e quindi non è possibile utilizzare alcuna analisi dall'esterno per riconoscere l'esistenza o meno di coscienza. Ancora: possiamo costruire una rete in grado di forzare un avvio di coscienza in un sistema artificiale? Attualmente la risposta è no, perché le macchine di calcolo oggi sviluppate hanno una configurazione hardware ad informazione integrata nulla, e quindi non possono produrre coscienza [15] . All'orizzonte tuttavia si profila la costruzione di macchine diverse, basate su altri principi fisici, con *hardware* neuromorfici, che potrebbero teoricamente essere in grado di produrre informazione integrata non nulla.
Ma, ammesso che ciò sia possibile, abbiamo davvero bisogno di questo?

Intelligenza artificiale, pensiero magico e cattura culturale

L'intervento di Daniela Tafani parte dalla constatazione che è sempre più diffusa la tendenza ad affidare a sistemi di apprendimento automatico (*machine learning)* decisioni che possono produrre effetti significativi sulla vita delle persone. Si va dalla valutazione della probabilità di rischio di recidiva di reati, alla selezione del personale (*recruiting*), così come alla valutazione di affidabilità

economica nel caso di prestiti bancari o altro ancora. Allineandosi inevitabilmente a quanto già avviene nelle attuali strutture sociali, gli strumenti di IA, che basano le loro capacità di risposta su dati grezzi raccolti senza filtro dalle società tecnologiche, ne riproducono le discriminazioni: verso persone di colore, donne e poveri, significativamente, ma comunque anche verso qualunque gruppo di persone statisticamente percepito dagli algoritmi come minoranza. Gli strumenti statistici che fanno da base dell'apprendimento automatico macinano enormi quantità di dati, ed è proprio la natura non filtrata e puramente statistica delle operazioni che conduce a discriminazioni verso tutti coloro che risultano marginali in termini probabilistici. In altre parole verso coloro che non rappresentano la grande maggioranza dei casi presenti nelle nostre attuali società, generalmente bianche, a predominio maschile, ricche, occidentali. Già questo dovrebbe esser sufficiente a mettere in discussione la percezione – errata – di oggettività dell'algoritmo, percezione che invece tende a persistere, espandendosi fuori da una qualsivoglia logica razionale. Un risultato che proviene dalla IA tende ad essere vero a prescindere, trascurando del tutto le modalità reali che sottendono alla produzione del risultato. Sei povero perché vivi in un quartiere degradato, oppure vivi in un quartiere degradato perché sei povero? Qual è la giusta correlazione? È sufficiente essere povero e vivere in un quartiere degradato per giudicarti e farti rimanere in galera in quanto una IA ti ritiene potenzialmente recidivo a commettere reati? I sistemi di IA, in questi casi, sono strumenti che producono risultati con effetti nel mondo reale, del tutto al di fuori di ogni trasparenza di funzionamento. Il loro utilizzo predittivo difficilmente può portare a risultati diversi da altre modalità più o meno casuali di decisione, o magari anche da valutazioni di natura astrologica. E quand'anche si riuscisse a far accettare la impossibilità di ottenere previsioni certe dalla tecnologia presente, ci pensa l'immaginario indotto a sostanziare gli avvenimenti, con valutazioni tirate via dritte dritte dalla fantascienza. Infine a chiudere il cerchio arriva l'ineluttabilità dell'avanzamento tecnologico, che magicamente cresce su se stesso passo dopo passo, come le scimmie che prima o poi arriveranno alla luna arrampicandosi su alberi sempre più alti e su se stesse. Riferendosi all'utilizzo della IA in qualità predittive (sarebbe meglio dire oracolari) è stato coniato il termine "olio di serpente"[16], sulla falsariga delle pozioni magiche, atte a risolvere i più svariati problemi, vendute a suo tempo nel *Far west* degli Stati Uniti.

Tutto ciò produce danni. È sulla base di questi assunti, ad esempio, che le cosiddette auto a guida autonoma continuano a circolare negli USA, nonostante che più o meno frequentemente falcidino persone che si muovono in modalità non riconosciute dall'algoritmo. Ma non è mai la guida autonoma a sbagliare, in quanto il gestore del sistema può facilmente provvedere ad interromperla un secondo prima degli impatti.

Quello che viene trascurato nella diffusione massiva degli strumenti di IA è che nella conoscenza umana esiste il senso comune, conoscenza condivisa e non codificata che nessuno strumento artificiale ad oggi possiede. Ed è questa conoscenza che impedisce di consentire a un bambino che gioca di dare fuoco alla casa!

Tra le narrazioni sulla IA, date per scontate e che condizionano qualsiasi discorso su di essa, esiste il principio della *inevitabilità della tecnologia*, principio che genialmente è in grado di fare evaporare le responsabilità. Nonostante OpenAI, creatore di Chat-GPT, nella sua pagina di presentazione dichiari di voler andare verso una IA generale, quelli che oggi esistono sono sempre e solo sistemi di IA ristretta, in grado di eseguire uno o pochi compiti specifici, a patto che quello che incontrano non sia eccessivamente diverso dai dati di partenza. I test che Chat-GPT supera provengono dalla base dati che le è stata data in pasto, e infatti cambiando le domande, le risposte spesso sono errate come errati sono, almeno al momento, i risultati di moltiplicazioni a più cifre. In questo senso la IA è un artefatto come un altro, prodotto da persone, e le modalità con cui tali artefatti vengono proposti ed utilizzati sottendono esercizi del potere, evidenziando in tal senso caratteristiche che spesso implicano precise strategie di *policy*.
Al contrario dell'enfatizzazione corrente, la domanda da fare sarebbe: perché questi sistemi sono chiusi? Perché non sono trasparenti?

Negli Stati Uniti, la *Federal Trade Commission* ha denunciato la diffusa pratica aziendale di utilizzare l'espressione "intelligenza artificiale" quale "termine di *marketing*"[17]. Nonostante queste evidenti criticità queste macchine vengono vieppiù adottate all'interno di sistemi di *governance* politica per prendere decisioni di carattere socioeconomico; con il risultato di

arrivare al punto che la decisione presa dalla IA, resa norma sottratta a chi avrebbe dovuto realmente decidere (siano essi funzionari, giudici o politici), produce ciò che pretende di prevedere. Quando questo si verifica, gli effetti possono essere rilevanti per la vita delle persone, sottoposte ad un giudizio definito oggettivo ma che tale non è. Di fatto viene a crearsi una presa di posizione, di natura pienamente politica, a favore della perpetuazione dello *status quo*.

I sistemi di apprendimento automatico sono sistemi resi possibili dalle grandi quantità di dati e dalla enorme potenza di calcolo in possesso delle grandi aziende tecnologiche. Lo sviluppo del loro modello di *business*, basato sul capitalismo della sorveglianza[18], la raccolta massiva ed indiscriminata di dati utilizzati per generare la vendita agli inserzionisti di una profilazione individuale a fini commerciali, ha permesso loro di avviare lo sviluppo dei *Large Language Models.* Non agendo su modelli di comprensione, nonostante che larvatamente lo si lasci intendere, ma su modelli statistici, è stato possibile costruire i progressi sostanziali visibili in determinati campi specifici (si pensi ad esempio al riconoscimento del linguaggio), ma ciò contemporaneamente ha dato alle *Big tech* la possibilità di espansione illimitata verso settori di intervento fino a poco tempo fa pensati impossibili. I prodotti che circolano, non solo ci riconoscono nei nostri comportamenti, ma ci giudicano, valutandoci positivamente o negativamente secondo la nostra collocazione nelle distribuzioni statistiche. Le *Big tech* hanno finanziato e finanziano università e centri di ricerca, dove era ben chiaro fin dall'inizio la fallacia dei sistemi di intelligenza artificiale nell'utilizzo in campi sensibili come quelli predittivi. La scelta è stata quella di minimizzare il problema, spingendo la penetrazione nel mercato di strumenti predittivi non costruibili su basi logiche, ma solo su correlazioni statistiche non filtrate in alcun modo. Il problema delle discriminazioni non è risolvibile costruendo sistemi che si basano sulla raccolta massiva ed indiscriminata dei dati, seppur depurati a valle dei contenuti più marcatamente inaccettabili. Ed è questa impossibilità che viene negata dalle *Big tech*, che operano diffondendo l'idea di una correggibilità dei sistemi basata su una presunta etica della IA, gestita da loro stesse e comunque posta al di fuori di ogni quadro giuridico proprio dello Stato di diritto.

Nei fatti si forza verso un predominio assoluto dell'algoritmo, tendendo ad escludere ogni quadro normativo preesistente. Con il risultato globale di avere un oggetto sempre più surrettiziamente antromorfizzato cui gli umani non avrebbero altra alternativa che adattarsi.

Note e commenti in calce

1. In effetti Chat-GPT ha raggiunto, in soli 9 mesi dal lancio, risultati che possiamo definire eccezionali: le hanno misurato un quoziente di intelligenza pari a 155 (il valor medio è 100, ad Einstein viene attribuito un valore pari a 160), ha superato decine di esami professionali ed accademici, aiutato a scrivere verdetti e leggi, fatto diagnosi mediche, investito in borsa, dimostrato di possedere capacità finora pensate non possibili per una macchina.

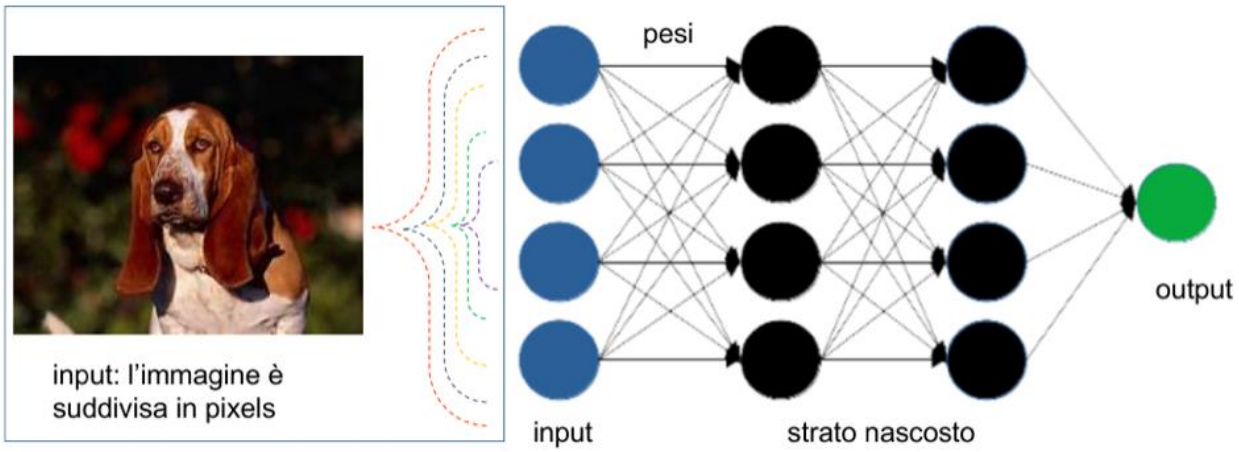
2. E' interessante qui riportare (in corsivo) le risposte che Chat-GPT (3.5) fornisce alle domande su intelligenza, consapevolezza e coscienza. "L'intelligenza è la capacità di apprendere, comprendere, elaborare informazioni e adattarsi a situazioni nuove o complesse. Si tratta della capacità di risolvere problemi, prendere decisioni informate, ragionare logicamente, apprendere da esperienze passate e applicare conoscenze in modi creativi. L'intelligenza può manifestarsi in molte forme diverse, comprese quelle umane e artificiali". In relazione alla propria intelligenza afferma "No, non sono intelligente nel senso in cui è definita l'intelligenza nelle persone. Sono un programma di intelligenza artificiale basato su modelli linguistici e dati di addestramento, e la mia "intelligenza" deriva dalla capacità di elaborare e generare testo in risposta a input basati su linguaggio naturale. Non ho una comprensione o una coscienza del mondo come un essere umano, e la mia "intelligenza" è limitata alla manipolazione del testo e alla generazione di risposte basate su modelli statistici e dati pregressi. Quindi, la mia "intelligenza" è molto diversa da quella umana e non implica una comprensione reale o una coscienza del mondo". Nello specifico dei Large Language Models tutte le elaborazioni sono basate su processi statistici, e non esiste allo stato nessuna consapevolezza reale della risposta data dalla macchina. Sulla coscienza "La coscienza è una qualità mentale che implica la consapevolezza di sé e dell'ambiente circostante. È la capacità di avere esperienze soggettive, pensieri, emozioni e sensazioni. La coscienza implica la percezione e la comprensione del proprio stato interiore e dell'ambiente esterno. È una caratteristica distintiva dell'essere umano, ma la natura della coscienza è ancora oggetto di dibattito e ricerca tra scienziati, filosofi e teologi. La coscienza umana è spesso associata alla riflessione, all'autocoscienza e alla capacità di attribuire significato alle esperienze". Le

IA non hanno esperienze soggettive, emozioni o percezioni. Generano risposte in base a modelli statistici, ma non comprendono veramente il significato delle informazioni che elaborano. Inoltre le IA sono strumenti neutri che rispondono alle istruzioni e ai dati di input. Non hanno desideri, intenzioni o preferenze. Possono sembrare intelligenti nelle loro risposte, ma questa intelligenza è il risultato di algoritmi e dati, non di una mente consapevole.

3. In altre parole la domanda riguarda la possibilità della macchina di diventare senziente e cosciente, ed in ultima analisi capace di costruire proprie visioni delle cose, e sulla base di queste prendere decisioni in maniera autonoma. La coscienza di sé, in quanto esperienza fenomenologica soggettiva, secondo alcune visioni farebbe emergere anche questioni etiche relative a diritti e doveri nei confronti della macchina. Più pericolosamente, come paventato in una lettera appello di richiesta di moratoria nello sviluppo della IA (https://futureoflife.org/open-letter/pause-giant-ai-experiments/), una IA costruita con volontà malevoli potrebbe far correre all'intero genere umano rischi di estinzione. Al contrario, in una visione da incubo trans-umanista diffusa tra i miliardari dell'High Tech (alcuni dei quali peraltro firmatari anche dell'appello di moratoria) sarà l'incorporazione degli umani in macchine intelligenti a permettere l'espansione della civiltà (naturalmente capitalistica) e la colonizzazione dell'universo (vedere ad esempio W. Mac Askill, What we owe the future, Basic Book, New York 2022).

4. Tecnologie di questo tipo sono alla base dei sistemi di videosorveglianza e dei sistemi di riconoscimento facciale.

5. Mutuando la terminologia del cervello biologico, una rete neurale informatica è costituita da sinapsi e neuroni. Un modello molto semplice di rete neurale è rappresentato nella figura sotto riportata. Tale modello, detto "rete di soglie binarie direzionali" (W. S. McCullog and W.H.Pitts, A logical calculus of the ideas immannet in nervous activity, Bullettin of mathematical biophysics, 5, 115-133 (1943.), rappresenta una rete neurale nella quale i nodi sono sono i neuroni e le connessioni le sinapsi.



Nel riconoscimento dell'immagine questa viene scomposta nei singoli pixel, i quali costituiscono i valori di ingresso della rete neurale digitale. I nodi della rete filtrano i dati attraverso l'assegnazione di una "importanza differenziata" a ciascuno di essi . In altre parole ogni dato in ingresso su ogni nodo viene pesato. In uscita dall'insieme dei nodi del primo strato, il dato viene ulteriormente valutato nella sua importanza, permettendo o meno il suo passaggio verso lo strato successivo. Questo fino ad arrivare all'output finale, che in questo caso è "cane". Il modello è stato via via ottimizzato, introducendo il possibile cambiamento dei pesi (D. O. Hebb, The organization of behavior; a neuropsychological theory. Wiley, New York, 1949), l'orientamento mirato verso gli obiettivi aspettati (F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386–408 (1958)), nonché meccanismi di automazione nel cambiamento dei pesi detti backpropagation. La backpropagation parte da un assunto di risultato, che in sé può essere giusto o sbagliato (ad esempio: la figura rappresenta un cane o un ladro carponi?). In fase di addestramento se il risultato è esatto si rafforzano i pesi dei dati passanti nelle connessioni che lo hanno prodotto, se è sbagliato se ne diminuisce il valore. Le potenze di calcolo attuali permettono di avere strutture neurali costituite da moltissimi strati e nodi, con valori che raggiungono un ordine di grandezza simile a quelli presenti nel cervello umano, valutabile in 100 miliardi di nodi e 1 milione di miliardi di connessioni.

6. In generale un linguaggio è formato da lettere che si raggruppano tra loro a formare parole, e parole che si raggruppano a formare frasi. Le modalità con cui lettere e parole si raggruppano possono essere valutate su base probabilistica. I sistemi di autoapprendimento delle reti neurali utilizzano l'enorme insieme di dati presenti in Internet per associare parole e produrre frasi che abbiano statisticamente un senso compiuto. Fornita una frase in input ad un LLM questo ne determina il senso statistico e produce una risposta che sia

sia statisticamente correlata. Il grado di apprendimento del modello, che sostanzialmente deriva dal volume dell'insieme di dati a cui attinge, determina la bontà della risposta.

7. L'encoder cattura il significato di ogni parola o simbolo e lo converte in un formato comprensibile per la macchina, lo elabora e lo passa alla successiva elaborazione da parte del decoder.

8. Il decoder prende la rappresentazione dell'input che l'encoder ha generato e la utilizza per generare l'output parlando in modo coerente attraverso la generazione di una sequenza di parole o simboli che abbia senso in base al contesto.

9. L'autoattenzione (A. Vaswani, N. Shazeer,N. Parmar,J. Uszkoreit, L. Jones, A. N. Gomez, I. Polosukhin. (2017). Attention is all you need, https://doi.org/10.48550/arxiv.1706.03762) consente al modello di linguaggiodi dare maggiore importanza a certe parole o parti del testo in base a ciò che sta cercando di codificare o decodificare. Nei cosiddetti transformer sia l'encoder che il decoder contengono meccanismi di autoattenzione.

10. Una logica del tutto simile opera nella comprensione di una domanda e nella successiva articolazione della risposta. Chat-GPT, prodotta da OpenAI, è al momento senz'altro il più noto e sviluppato LLM. Il termine GPT sta per "Generative Pre-trained Transformer": si tratta di un transformer preaddestrato (sull'enorme base dati di internet) generativo, cioè in grado di produrre (generare) risposte.

11. In una rete biologica il neurone è l'unità cellulare che costituisce il tessuto nervoso. Grazie alle sue peculiari proprietà fisiologiche e chimiche è in grado di ricevere, elaborare e trasmettere impulsi nervosi sia eccitatori che inibitori. Esiste una ulteriore differenziazione del neurone biologico rispetto a quello artificiale. Questa consiste nella suddivisione tra una parte adibita a trasmettere un segnale in uscita (detta assone), e una parte attraverso la quale ogni cellula nervosa riceve comunicazione dalle altre cellule nervose (detta dendrite), L'estensione degli assoni può essere molto elevata e non limitata ai nodi limitrofi. Ogni volta che un neurone fa passare la comunicazione per una sinapsi, l'informazione che passa cambia il suo peso tra i due nodi adiacenti ma non cambia mai il segno. Infine, le reti neurali biologiche sono dotate di connettività rientrante (meccanismo che, tra l'altro, consente il coordinamento reciproco delle diverse aree cerebrali per far nascere nuove funzioni), adattabilità attraverso meccanismi di riconnessione, sostituzione delle giunzioni sinaptiche, oltre che di una importante correlazione ad ampio raggio, in grado di mutare nello spazio e nel tempo. Il comportamento delle reti neurali biologiche è dinamicamente mutevole (plastico) anche in assenza di supervisione esterna. Questo, assente nelle reti informatiche, è alla base dei processi di apprendimento.

12. G.Tononi, PHI. Un viaggio dal cervello all'anima, Codice edizioni, Torino, 2014.

13. Solo le relazioni che si verificano all'interno della rete attraverso connessioni rientranti producono una quantità di informazione integrata non nulla. Nelle reti puramente unidirezionali (feed forward) non esistono connessioni rientranti in grado di produrre informazione integrata non nulla. Tutta l'informazione è prodotta attraverso meccanismi di training e sotto supervisione.

14. E.M. Bender, T. Gebry, A. McMillan-Major, S. Shmitchell. On the Dangers of Stochastic Parrots: CanLanguage Models Be Too Big?" FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March2021Pages 610–623, https://doi.org/10.1145/3442188.3445922. A seguito della diffusione dell'articolo Timnit Gebru e Margaret Mitchell (autrice con lo pseudonimo Shmargareth Shmitchell) sono state licenziate o indotte a licenziarsi da Google. Vedere anche il precedente nostro articolo Chat-GPT e le alre: paura del futuro distopico?, https://sbilanciamoci.info/chatgtp-e-le-altre-paura-del-futuro-distopico/)

15. La teoria dell'informazione integrata non è l'unica teoria che cerca di esplicitare le modalità attraverso le quali può emergere una coscienza. Esistono altri approcci di diversa natura. La teoria dell'elaborazione ricorrente ipotizza che il passaggio delle informazioni attraverso cicli di feedback sia la chiave per la coscienza. La teoria dello spazio di lavoro neuronale globale sostiene che la coscienza nasce quando flussi indipendenti di informazioni passano attraverso un "collo di bottiglia" per combinarsi tra loro. Teorie di ordine superiore suggeriscono che la coscienza implichi un processo di rappresentazione e annotazione degli input di base ricevuti dai sensi. Altre teorie sottolineano l'importanza dei meccanismi di controllo dell'attenzione e la necessità di un corpo che riceva feedback dal mondo esterno. Questi menzionati sono tutti approcci che si muovono all'interno di una visione meccanicistica della realtà, mutuata dalla fisica classica.
Qui vogliamo anche citare approcci di tipo non meccanicistico, quali quelli del premio Nobel Penrose (S. Hameroff, R. Penrose. 2014. Consciousness in the Universe, Physics of Life Reviews 11, pag. 39–78) e di Federico Faggin inventore del microprocessore (F. Faggin,. Irriducibile – La coscienza, la vita, i computer e la nostra natura, Mondadori, 2022), che ipotizzano che all'origine dei fenomeni di coscienza vi siano fenomeni quantistici o quanto-relativistici. Con eventuali computer quantistici si supererà questa barriera? Se la coscienza è un fenomeno quantistico sarebbe soggetta anche ai limiti conoscitivi imposti dalla meccanica quantistica (correlati al principio di indeterminazione di Heisenberg): detto in maniera grezza un sistema

quantistico è la sovrapposizione di più stati, nel momento in cui lo si indaga (ovvero si compiono osservazioni e misure su di esso) il sistema "collassa" in un singolo stato, che è quello che viene osservato, ma non è la situazione del sistema prima della misura. Anche qui ci troviamo di fronte a un limite invalicabile, che poi si formalizza in un "teorema di impossibilità" – il teorema di Bell, che afferma essere impossibile che la meccanica quantistica sia una teoria ad un tempo "completa" e "locale", cioè che non ammetta "magiche azioni a distanza" come l'entanglement (fenomeno che per decenni ha destato perplessità nei fisici "ortodossi", e la cui verifica sperimentale ha valso il premio Nobel nel 2022 a A. Aspect, J.F. Clauser e A. Zeinlinger, i tre fisici che lo hanno verificato). Scriviamo "anche" perché, oltre quello di Bell, anche i teoremi di Gödel, di cui ci siamo occupati nel precedente articolo su Sbilanciamoci, sono teoremi di impossibilità. Insomma, Gödel e Bell sono le due facce di una medaglia che segnala i limiti che abbiamo nella nostra comprensione della natura.

16. Si veda ad esempio A. Kumar, Automation of Data Prep, ML, and Data Science: New Cure or Snake Oil? ,SIGMOD '21: Proceedings of the 2021 International Conference on Management of DataJune 2021Pages 2878–2880, https://doi.org/10.1145/3448016.3457537.

17. Si veda https://www.ftc.gov/system/files/ftc_gov/pdf/remarks-of-chair-lina-m-khan-re-joint-interagency-statement-on-ai.pdf, consultato il 18 ottobre 2023.

18. S.Zuboff, Il capitalismo della sorveglianza, Luiss University Press, Roma, 2019.

Sbilanciamoci!, 19 ottobre 2023