INTELLIGENZE ARTIFICIALI GENERATIVE E MEDIAZIONE INFORMATIVA: UNA INTRODUZIONE

GINO RONCAGLIA
Dipartimento di Filosofia, comunicazione e spettacolo,
Università degli studi di Roma Tre
gino.roncaglia@uniroma3.it

Biblioteche oggi Trends, vol. 9, n. 1 (giugno 2023)
p. 13-26,DOI: 10.3302/2421-3810-202301-013-1
ISSN: 2421-3810

Premessa

Le intelligenze artificiali generative sono da diversi mesi al centro di una notevole attenzione mediatica, ed è abbastanza diffusa la previsione che il loro sviluppo porterà rapidamente a cambiamenti anche radicali in molti ambiti professionali, incluso il mondo della mediazione informativa.
Si tratta di una previsione giustificata? E quali sono gli sviluppi che possiamo aspettarci al riguardo? Per discutere questi temi, che saranno ripresi in diversi fra gli interventi raccolti in questo fascicolo di *Biblioteche oggi Trends*, occorre per prima cosa capire di cosa esattamente stiamo parlando: cosa sono, e come funzionano, le intelligenze artificiali generative? In questo articolo cercherò di presentare – in forma necessariamente sintetica – il contesto all'interno del quale si è sviluppato il lavoro su questi sistemi, i principi di funzionamento e le caratteristiche di alcuni di essi (in particolare di quelli basati sulla generazione di testi attraverso *transformer*, come GPT e ChatGPT), i principali problemi riscontrati e una prima, assai parziale riflessione sull'impatto che potranno avere nel futuro[1].

Il contesto: IA e reti neurali

Anche se la riflessione sulla possibilità di costruire macchine 'intelligenti' (in un qualche senso del termine) è molto più antica[2], il lavoro sull'intelligenza artificiale[3], collegato agli sviluppi nel campo dell'informatica e alla rivoluzione digitale, inizia negli anni Cinquanta del secolo scorso ed è legato soprattutto a due nomi e due occasioni che hanno contribuito in maniera determinante a delinearne l'impostazione iniziale: quello di Alan Turing, che nell'articolo del 1950 *Computer machinery and intelligence* [Turing, 1950] ha posto le basi teoriche della riflessione sul rapporto fra intelligenza artificiale e intelligenza umana, e quello di John McCarthy, che ha organizzato il fondamentale seminario svoltosi nell'estate 1956 al Dartmouth College. È nel documento preparatorio di tale incontro che compare l'espressione *artificial intelli-*

Per tutti i siti web la data di ultima consultazione è il 4 maggio 2023.

[1] Sono grato a Simone Arcagni, Fabio Ciotti, Fabio Ciracì, Lauro Colasanti, Cesare Cozzo, Mario De Caro, Derrick de Kerckhove, Grazia Farina, Maurizio Lana, Francesco Leonetti, Giorgio Parisi, e ai revisori anonimi di *Biblioteche oggi Trends* per osservazioni, commenti, critiche e suggerimenti sui contenuti di questo articolo; la responsabilità di errori e imprecisioni resta ovviamente solo mia.
[2] Il bel volume [Cave - Dihal - Dillon, 2020] fornisce nella sua prima parte numerosi esempi affascinanti, dagli automi ed esseri artificiali intelligenti presenti nelle narrazioni omeriche alla diffusa – e ovviamente falsa – leggenda tardo-medievale e rinascimentale dell'androide o della testa di androide intelligente la cui costruzione era attribuita ad Alberto Magno e la cui distruzione (volontaria o involontaria) era da alcuni attribuita al più noto fra gli allievi di Alberto, Tommaso d'Aquino.
[3] Suggerire un testo che possa fornire insieme una introduzione e uno strumento di riferimento generale sul tema dell'intelligenza artificiale è compito non banale, ma la risposta probabilmente migliore è rappresentata dal classico [Norvig - Russell, 2021], manuale utilizzato da generazioni di studenti – tanto da essere spesso citato semplicemente con l'acronimo AIMA – e arrivato ormai alla quarta edizione (la prima è del 1995).

gence [McCarthy *et al.*, 1955], ed è in questo contesto che nasce l'indirizzo della cosiddetta 'intelligenza artificiale forte': l'idea che gli sviluppi dell'informatica possano permettere di creare macchine dotate di un'intelligenza simile alla nostra. HAL 9000, il computer al centro del film di Stanley Kubrick *2001: A Space Odyssey*, è la rappresentazione forse più nota di questa ambizione. Fra i consulenti di Kubrick per quel film, uscito nel 1968, vi era Marvin Minsky, che aveva affiancato McCarthy nella creazione del laboratorio di intelligenza artificiale del MIT, e HAL rappresenta per molti versi il prototipo di 'computer intelligente' che negli anni Sessanta del secolo scorso sembrava quasi a portata di mano: basti pensare che meno di 35 anni separavano la data di realizzazione del film da quella della sua ambientazione, e che ancor più breve era l'orizzonte temporale previsto dal premio Nobel Herbert Simon dopo la sua partecipazione al seminario di Dartmouth: appena vent'anni [Hoffmann, 2022a].

Quello che, col senno di poi, potrebbe sembrare eccessivo ottimismo sulla rapidità dello sviluppo tecnologico era in realtà il risultato di una doppia assunzione sulla natura dell'intelligenza umana: da un lato, l'idea che la nostra intelligenza sia in primo luogo linguistica (manifestiamo la nostra intelligenza in primo luogo attraverso l'uso del linguaggio); dall'altro, l'idea che il linguaggio possa essere analizzato in termini di sistema governato da regole, e che queste regole siano in linea di principio formulabili in maniera tanto precisa e rigorosa da poter essere programmate e utilizzate da un sistema informatico. Due assunzioni con una lunga storia alle spalle (basti pensare all'idea del linguaggio come calcolo in Hobbes e alle intuizioni di Leibniz sulla convergenza del calcolo logico e del calcolo binario in direzione di quella che potremmo chiamare 'computabilità del mondo') ma legate in primo luogo, rispettivamente, alla *linguistic turn*, la 'svolta linguistica' propria della prima filosofia analitica, che poneva l'uso del linguaggio al centro della riflessione filosofica, e agli sviluppi della logica formale, a sua volta fra i tratti distintivi del panorama culturale novecentesco. Si aggiunga, sul secondo fronte, il lavoro che proprio in quel periodo portavano avanti linguisti come Noam Chomsky, che ipotizzavano l'esistenza di una struttura grammaticale profonda comune alle diverse lingue e a sua volta analizzabile in termini rigorosamente formali.

Se la nostra intelligenza è soprattutto linguistica, e se il linguaggio può essere considerato come un sistema governato da regole precise e formalizzabili, l'idea di un computer programmato attraverso quelle regole in modo da saper usare il linguaggio non appare più troppo peregrina. Nel prospettarla, Turing aggiunge un ulteriore, importante tassello a questo quadro: il criterio per attribuire intelligenza a un essere diverso da noi può essere solo il suo comportamento intelligente. Attribuiamo intelligenza alle altre persone basandoci sul loro comportamento, senza bisogno di andare a verificare ogni volta che la fisiologia e il funzionamento del loro cervello siano effettivamente analoghi a quelli del nostro. Un criterio analogo, argomenta Turing, dovrebbe essere applicato alle macchine intelligenti: se il loro comportamento intelligente (e di nuovo, in particolare, il loro comportamento 'linguistico' intelligente) è analogo al nostro, non abbiamo ragione per non attribuire loro un'intelligenza analoga alla nostra, anche se alla base c'è un sistema informatico evidentemente diverso dal nostro cervello biologico.

È questa impostazione che giustifica l'idea del test di Turing[4] come strumento per attribuire intelligenza a una macchina: se un computer è in grado di ingannare un esaminatore attraverso un comportamento (linguistico) intelligente indistinguibile da quello umano, non vi è ragione per non considerarlo intelligente.

A partire dagli anni Settanta del secolo scorso, tuttavia, tanto le due premesse rappresentate dall'idea della natura prevalentemente linguistica dell'intelligenza e dall'idea che il linguaggio abbia una struttura profonda analizzabile in termini rigorosamente formali e dunque riproducibile attraverso algoritmi, quanto l'assunzione che un comportamento intelligente sia un criterio sufficiente all'attribuzione di intelligenza, sono state messe in discussione. Per un verso, sottolineando l'importanza di dimensioni non linguistiche dell'intelligenza (legate ad esempio all'orientamento nello spazio e in generale al ruolo della nostra dimensione corporea, all'intelligenza emotiva, all'esistenza di forme di intelligenza basate su componenti simboliche non linguistiche, all'esistenza di motivazioni all'azione intelligente non sempre espresse o esprimibili in forma linguistica o antecedenti rispetto alla loro espressione linguistica ecc.) ed evidenziando i numerosi problemi presenti nelle varie proposte di modelli formali del linguaggio, nessuno dei quali appare di fatto in grado di dar conto in maniera soddisfacente del nostro comportamento linguistico e di farlo riprodurre da una macchina. Per altri versi, attraverso una critica

[4] Nel test di Turing un computer e un essere umano comunicano indipendentemente con un esaminatore esterno attraverso un apparato per la trasmissione di testi (ad esempio una telescrivente). Se, dopo una interazione sufficientemente lunga e articolata, l'esaminatore non riesce a identificare correttamente quale sia l'output del computer e quale sia quello umano, secondo Turing possiamo attribuire al computer una forma di intelligenza produttiva funzionalmente analoga alla nostra. Ho discusso in maniera più approfondita il tema, e fornito alcune essenziali indicazioni bibliografiche al riguardo, in [Roncaglia, 2014].

dell'idea che un comportamento apparentemente intelligente sia sufficiente ad attribuire intelligenza a una macchina (l'argomento della stanza cinese[5] proposto dal filosofo statunitense John Searle [Searle, 1980] rappresenta una critica al test di Turing proprio da questo punto di vista).

Anche dal punto di vista strettamente tecnologico, peraltro, il lavoro nel campo dell'intelligenza artificiale forte non ha portato ai rapidi sviluppi che erano stati ipotizzati. Di fatto, già intorno alla metà degli anni Settanta del secolo scorso molti ricercatori – incluso lo stesso Minsky – hanno cominciato ad abbandonare questa prospettiva, a favore di un ventaglio molto più ampio ma meno ambizioso di indirizzi di ricerca, spesso riuniti nella categoria-ombrello di 'intelligenza artificiale debole'. Fra di essi, il lavoro su ambiti più ristretti e specifici (è la strada, ad esempio, dei sistemi esperti settoriali, che almeno inizialmente adottavano strumenti di modellizzazione dei processi di inferenza logica elaborati nell'ambito dell'IA forte, applicandoli però a campi assai più limitati: quello medico, quello giuridico ecc.); il tentativo di riprodurre comportamenti intelligenti di organismi meno complessi dell'uomo (così, ad esempio, negli anni Ottanta Rodney Brooks lavorava sulla realizzazione di piccoli robot che avrebbero dovuto manifestare un livello di intelligenza paragonabile a quello degli insetti); il lavoro di automazione di processi o dispositivi dedicati a scopi particolari (un esempio recente è rappresentato dalle macchine a guida autonoma) e così via.

La reti neurali, che esplorano la possibilità di replicare attraverso reti di neuroni artificiali alcuni aspetti del funzionamento del nostro cervello, erano state proposte come uno dei possibili ambiti di lavoro dell'intelligenza artificiale già nel documento preparatorio del seminario di Dartmouth [McCarthy *et al.*, 1955], e rappresentavano in sostanza una ibridazione fra l'intelligenza artificiale e le riflessioni sulla relazione fra uomo e macchina proposte da un altro degli indirizzi di ricerca dell'epoca, la cibernetica[6]. Ma almeno all'inizio sembrava difficile ipotizzare la creazione di reti neurali abbastanza potenti da svolgere compiti 'intelligenti' realmente complessi: un'idea rafforzata dalle conclusioni di un testo di Minsky e Papert di cui parleremo fra un attimo. In un certo senso, dunque, negli ultimi decenni del secolo scorso anche le reti neurali finirono nel calderone dei vari indirizzi di ricerca riuniti sotto l'etichetta di intelligenza artificiale debole. La situazione, però, era destinata a cambiare presto. Per capire il perché, occorre prima presentare sinteticamente la loro storia e alcune fra le idee chiave che ne hanno segnato lo sviluppo.

L'idea delle reti neurali nasce sulla base delle ricerche svolte già negli anni Quaranta del secolo scorso dal neurofisiologo Warren McCulloch in collaborazione con Walter Pitts, logico e matematico. McCulloch e Pitts proponevano [McCulloch - Pitts, 1943] di guardare al neurone come a una sorta di macchina di computazione, che sulla base delle informazioni raccolte dai 'dendriti' (piccole ramificazioni che permettono al 'soma' o nucleo di un neurone di ricevere stimoli dall'esterno) e processate dal soma produce un output trasmesso dall"assone' (il principale collegamento in uscita dal neurone) e dalle 'sinapsi' che da esso si diramano verso altri neuroni.

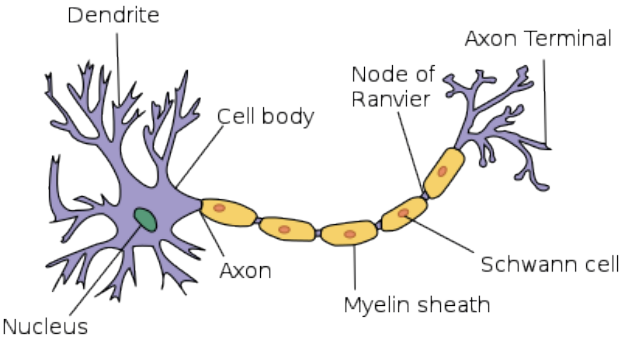

Fig. 1: Rappresentazione schematica di un neurone (fonte: <https://commons.wikimedia.org/wiki/File:Neuron.svg>)

Il lavoro computazionale del neurone viene rappresentato da McCulloch e Pitts attraverso una funzione che riceve come valori in ingresso gli input prodotti dai den-

[5] Searle immagina di essere chiuso in una stanza e di ricevere – con un meccanismo simile a quello del test di Turing – delle domande in cinese da un interlocutore esterno alla stanza. Searle non parla il cinese e dunque non capisce le domande ricevute, ma nella stanza ha a disposizione un manuale che associa simboli cinesi (fra cui quelli che costituiscono le domande ricevute) ad altri simboli cinesi (che costituiranno l'output). Può così individuare sul manuale i simboli ricevuti, copiare i corrispondenti simboli di output e trasmettere la risposta. L'esaminatore esterno avrà l'impressione che Searle, rispondendo correttamente in cinese, capisca il cinese, mentre in realtà non è così. Per Searle il computer che risponde al test di Turing compie la stessa operazione: produce meccanicamente un output seguendo un programma, ma non comprende le risposte che fornisce: non può dunque essere considerato intelligente. L'argomento di Searle è stato a sua volta variamente discusso e spesso criticato: Turing lo avrebbe probabilmente considerato come una variante dell'obiezione (4) considerata nel suo articolo [Turing, 1950], e avrebbe credo obiettato che se la stanza cinese è capace di rispondere adeguatamente a un numero indefinito di domande in cinese, è la stanza nel suo insieme (dunque non il solo Searle, ma l'insieme Searle + istruzioni, considerato come una sorta di 'black box') che 'parla' cinese.

[6] Sul rapporto fra la prima intelligenza artificiale e la cibernetica si vedano [Numerico, 2021, capitolo 1] e i numerosi riferimenti bibliografici ivi citati.

driti (ciascuno dei quali – in questo primo modello – può avere valore 0 o 1) e restituisce un output, a sua volta 0 o 1, in funzione degli input ricevuti. È facile capire che in questo modo si possono immaginare dei neuroni che si comportano come operatori logici (ad esempio un neurone AND restituirà il valore 1 se e solo se tutti gli input ricevuti hanno valore 1, un neurone OR restituirà il valore 1 se e solo se almeno uno degli input avrà valore 1, e così via). Dai nomi dei suoi due ideatori, questo modello computazionale è spesso chiamato 'neurone M-P'[7].

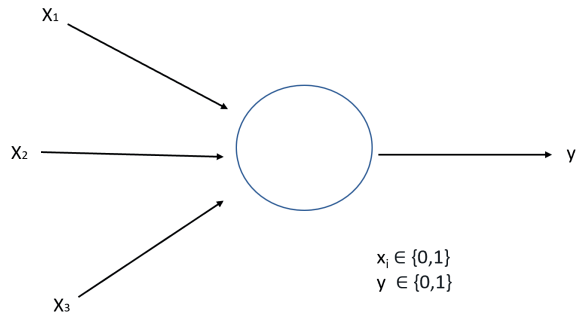

Fig. 2: Un esempio di neurone M-P con input e output binari. Il valore di ogni input X sarà 0 o 1 e il valore dell'output Y sarà anch'esso 0 o 1. Se vogliamo ad esempio che il neurone M-P corrisponda all'OR logico, Y avrà valore 1 se e solo se almeno uno degli X_i ha valore 1 (dunque, se la somma dei valori degli input è superiore a 0)

Alla luce delle nostre conoscenze attuali, quella proposta da McCulloch e Pitts è sicuramente una semplificazione estrema – influenzata dal logicismo un po' ingenuo dell'epoca – di una realtà assai più complessa; cosa di cui, del resto, gli stessi McCulloch e Pitts erano perfettamente consapevoli; quello che interessava loro era la costruzione di un modello matematico e logico, piuttosto che di un modello biologico accurato:

> The 1943 networks were only "possible" and "useful" – in no way did McCulloch and Pitts claim their model was a true description [of] known networks. McCulloch and Pitts acknowledged in their paper that their definition of a neuron was idealized, and that they made physical assumptions that were "most convenient for the calculus" (McCulloch & Pitts 1943, p. 116). Their method here was to begin with theoretical presuppositions and idealizations, and to construct hypothetical networks based on these presuppositions. As such, their diagrams represent hypothetical networks, formally equivalent to Boolean statements, and depict neurons that bear little resemblance to "real" neurons [Abraham, 2002, p. 21].

Questa semplificazione, tuttavia, rappresenta il primo e fondamentale passo di una nuova branca di ricerca, le neuroscienze computazionali, a loro volta cruciali per l'avvio del lavoro sulle reti neurali artificiali. Il passo successivo è costituito dal cosiddetto 'percettrone' (*perceptron*), la cui prima implementazione da parte di Frank Rosenblatt, psicologo e scienziato cognitivo statunitense, risale al 1958, e il cui modello è stato in seguito discusso e affinato – mostrandone tuttavia anche alcuni limiti[8] – nel 1969 da Marvin Minsky, che abbiamo già incontrato, e da Seymour Papert, un nome fondamentale non solo nel campo dell'intelligenza artificiale ma anche in quello delle teorie dell'educazione [Minsky - Papert, 1969]. Il percettrone accetta fra i suoi input non soltanto 0 e 1 ma qualunque valore reale; inoltre, è possibile assegnare un 'peso' diverso ai diversi input. Infine, è previsto un valore di soglia che determinerà o meno l'"attivazione' del percettrone (il percettrone attivato emette un output 1, il percettrone non attivato corrisponde a un output 0). In sostanza, il percettrone è attivato se e solo se la somma pesata dei valori di input raggiunge o supera il valore di soglia. Valore che a sua volta non è necessariamente fisso: possiamo considerare anch'esso come uno degli input della computazione effettuata dal percettrone (o meglio, come il peso – variabile – di un input arbitrario di valore 1). In questo modo, il valore di soglia potrà essere modificato o esplicitamente dal programmatore o – più frequentemente – in base all'input ricevuto da altri percettroni. La possibilità di modificare il valore di soglia e, più in generale, i pesi dei vari input durante l'addestramento permette al percettrone di 'imparare': le reti basate su percettroni – e, più in generale, le reti neurali – comprendono tipicamente uno strato di input, che riceve informazioni dal mondo esterno, (almeno) uno strato intermedio che elabora e 'apprende', e uno strato di output che restituisce all'esterno il risultato del lavoro della rete.
Se il percettrone prevede valori di input non limitati a 0 e 1 e ai quali possono essere attribuiti dei pesi, l'output resta però binario. Inoltre, il valore di soglia corrisponde a uno 'scalino' netto: l'output sarà sempre 1 se il valore di soglia è raggiunto o superato, anche di pochissimo, e sarà sempre 0 se non è raggiunto, anche se manca

[7] Per una discussione delle radici storiche del modello M-P e delle idee di McCulloch e Pitts si veda [Abraham, 2002]. Fra le molte presentazioni divulgative del modello suggerisco, anche per la chiarezza delle illustrazioni, [Chandra, 2018].
[8] Non li discuterò in questa sede, ricordando solo che il limite più evidente – l'impossibilità di gestire alcune tipologie di funzioni, fra cui la funzione logica XOR – è stato in seguito superato nelle reti neurali con più strati intermedi.

pochissimo a raggiungerlo. Il passo successivo è stato quello di sostituire allo scalino netto una – più realistica – soglia probabilistica, che corrisponde a una 'funzione di attivazione' (ad esempio una curva 'sigmoide'): il superamento del valore di soglia diventa così il momento in cui la probabilità di attivazione del neurone supera il 50%. I neuroni delle reti neurali artificiali utilizzate oggi funzionano in questo modo, permettendo – anche attraverso la scelta di funzioni di attivazione diverse – un addestramento ancor più fine dei singoli neuroni e della rete neurale nel suo complesso, e riducendone il determinismo.

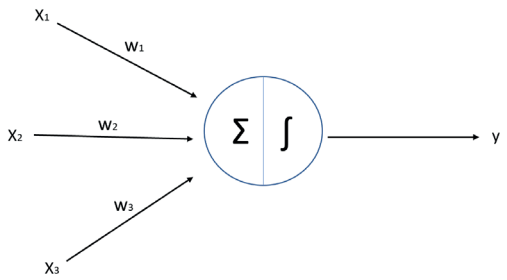

Fig. 3: Rappresentazione schematica di un neurone artificiale in una rete neurale: i valori X_1-X_n non sono necessariamente binari, a ogni input è applicato un peso W_1-W_n, e alla sommatoria pesata **Σ** dei valori in input è poi applicata una funzione di attivazione **∫**.

Tipicamente, una rete neurale viene addestrata sulla base del confronto fra l'output prodotto e quello desiderato, minimizzando progressivamente lo scostamento (o errore): un algoritmo di retropropagazione dell'errore (*backpropagation*) – sul cui funzionamento non mi soffermerò in questa sede – permette, attraverso meccanismi di iterazione, di modificare dinamicamente i pesi all'interno della rete in funzione della distanza fra il risultato ottenuto e quello atteso, ottenendo di fatto un apprendimento basato su *feedback* di rinforzo (quando la distanza dal risultato atteso diminuisce) o di indebolimento (quando invece aumenta). Il *feedback* può essere umano e/o automatico, e può essere anche affidato ad altre reti neurali. È importante notare che nelle reti neurali di oggi – estremamente complesse e il cui processo di addestramento prevede un lavoro assai 'costoso' in termini computazionali, sviluppato a partire da una grande quantità di informazioni di input – la situazione degli strati intermedi della rete è quasi completamente opaca: durante l'addestramento la rete modifica infatti pesi e valori in maniera autonoma, senza che chi la programma conosca effettivamente gli stati dei singoli neuroni che la compongono.

Questo vale in particolare quando gli strati o livelli intermedi della rete sono numerosi. In tal caso siamo davanti a quelle che vengono chiamate 'reti neurali profonde' (*deep neural network*). Gli strati intermedi consentono al sistema di costruire gerarchie di rappresentazioni dei dati, in cui gli strati più bassi apprendono caratteristiche di basso livello (come bordi o colori nelle immagini), mentre gli strati più alti apprendono caratteristiche di livello superiore (come oggetti o concetti). La capacità di apprendere rappresentazioni gerarchiche rende sistemi di questo tipo (i cosiddetti sistemi di *deep learning*[9]) particolarmente efficaci nell'affrontare problemi complessi, come il riconoscimento di immagini, la comprensione del linguaggio naturale e il riconoscimento vocale.

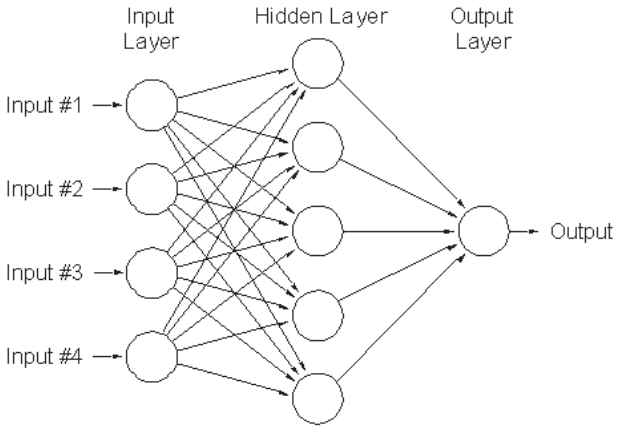

Fig. 4: Un esempio di semplice rete neurale (fonte: <https://commons.wikimedia.org/wiki/File:Neural_Network.gif>); normalmente, gli strati nascosti sono più di uno.

Come accennato, il lavoro sulle reti neurali si è mosso di fatto per alcuni decenni nell'ambito dell'intelligenza artificiale debole, con la realizzazione di reti neurali addestrate per scopi specifici (ad esempio il riconoscimento di forme o immagini fornito da sistemi di intelligenza artificiale 'discriminativa'). Tuttavia, le intelligenze artificiali generative di oggi – in particolare quelle legate alla generazione dei testi, come ChatGPT, e ancor più le sue varianti basate su agenti autonomi capaci di eseguire compiti complessi scomponendoli in più attività svolte in successione, come Auto-GPT e AgentGPT – stanno chiaramente superando almeno alcune di queste limitazioni, e sembrano ormai avvicinarsi all'idea di una intelligenza artificiale 'generalista', capace di rispondere a domande e di effettuare compiti relativi a uno spettro

[9] Fra le molte risorse esistenti, per una introduzione accessibile al *deep learning* suggerirei [Glassner, 2021].

assai ampio di situazioni e necessità diverse [Bubeck *et al.*, 2023]. Per certi versi, si tratta dunque di una sorta di rivincita del sogno dell'intelligenza artificiale forte, ripensato tuttavia in termini molto diversi rispetto al passato, e in cui la statistica, i *big data* e le associazioni probabilistiche all'interno di reti neurali profonde hanno sostituito la logica e la tradizionale programmazione deterministica come strumenti di riferimento.

La distinzione fra intelligenza artificiale forte e debole viene oggi spesso riformulata in termini di distinzione fra *artificial general intelligence* (AGI) e *narrow AI*, dove la AGI dovrebbe essere in grado di affrontare in maniera intelligente (almeno) le molte e diverse situazioni in cui gli esseri umani manifestano la loro intelligenza. A sua volta, la AGI è talvolta considerata come un passo verso forme di 'superintelligenza' (*superintelligence*)[10] superiori all'intelligenza umana. Riprendendo l'espressione proposta in [Kurzweil, 2005] si parla in questi casi di 'singolarità' (*technological singularity*) per denotare il momento in cui la crescita esplosiva della superintelligenza artificiale renderebbe di fatto superata l'intelligenza umana.

Non discuterò qui il tema – di grande interesse e per molti versi perturbante – dell'effettiva possibilità di costruire sistemi intelligenti che raggiungano il livello di una AGI o addirittura una forma di superintelligenza, limitandomi a notare come questa discussione sia ormai passata dall'ambito delle speculazioni quasi fantascientifiche alle pagine di libri, articoli e autori accademicamente rispettabili[11]. Questo naturalmente non implica che l'impresa sia poi davvero realizzabile, o che lo sia attraverso sistemi simili alle attuali intelligenze artificiali generative; in ogni caso, una discussione adeguata di questo tema richiederebbe una trattazione molto più estesa e competente di quella possibile in questa sede. Cercherò invece, sulla base del contesto fin qui delineato, di fornire qualche informazione in più sui sistemi di intelligenza artificiale generativa e sulle prospettive che essi aprono nel campo della mediazione informativa.

Le intelligenze artificiali generative

Basati su reti neurali profonde, i sistemi di intelligenza artificiale generativa costituiscono un sottoinsieme del *deep learning*, in cui l'obiettivo è produrre contenuti (che possono essere testuali, visivi, sonori, ma anche rappresentati da codice e programmi, giochi, ambienti virtuali, modelli 3D ecc.) in genere in risposta a un *prompt* (o richiesta) da parte dell'utente; *prompt* che sarà spesso – ma non necessariamente – testuale.

Così, ad esempio, i più noti sistemi di intelligenza artificiale generativa che producono immagini – ricordiamo, a solo titolo esemplificativo, Midjourney, Stable Diffusion, Dall-E ecc. – funzionano sulla base di *prompt* testuali che dovranno fornire una sorta di 'descrizione' a parole dell'immagine che si desidera generare e, analogamente, i più noti sistemi di intelligenza artificiale generativa che producono testi lo fanno in risposta a un *prompt* testuale.

Va ricordato comunque che non tutti i sistemi basati su reti neurali sono generativi: non mancano sistemi che hanno obiettivi diversi, ad esempio discriminare o classificare testi o immagini. E anche quando l'obiettivo è quello di generare un contenuto, ciò può avvenire senza partire necessariamente da un *prompt* testuale. Ad esempio, nel caso delle immagini l'obiettivo potrebbe essere quello, opposto, di generare una descrizione testuale partendo dall'analisi di un'immagine fornita come input (e un compito analogo potrebbe riguardare un video); oppure si potrebbe voler generare immagini combinando un *prompt* testuale e un'immagine fornita come esempio. In questo articolo mi soffermerò esclusivamente sulle intelligenze artificiali generative, e in particolare su quelle – come GPT o ChatGPT – che generano testi in risposta a *prompt* dell'utente.

Questi sistemi, sviluppati anche a partire dalle ricerche nel campo dell'elaborazione del linguaggio naturale (NLP, o *natural language processing*)[12], funzionano sempre partendo da un vasto corpus di testi, utilizzato per la costruzione del modello. Il primo passo è quello di selezionare il corpus e di prepararlo per l'analisi (*preprocessing*). Lo si fa attraverso la 'tokenizzazione', fase in cui il testo viene ripulito e suddiviso in *token*: unità più piccole che possono essere singole parole o morfemi di più basso livello, ma anche singoli caratteri o *n*-grammi (gruppi di *n* caratteri), a seconda del modello di tokenizzazione usato. Ad esempio, la parola 'dinosauro' potrebbe essere analizzata come 'dino-sauro' (ma anche, volendo, come 'dino-saur-o', o in altri modi ancora).

Segue la fase dell'apprendimento autonomo (*unsupervised learning*), durante la quale la rete neurale impara, sempre sulla base del corpus di partenza e aggiustando progressivamente i valori associati ai *token* e i pesi dei propri collegamenti interni, a predire il *token* successivo sulla base di quelli precedenti. È in questa fase che si crea

[10] Il testo di riferimento (molto discusso) sul tema è [Bostrom, 2014].
[11] Una utile rassegna critica è in [Hoffmann, 2022a]; per una trattazione più ampia, cfr. [Hoffmann, 2022b]. Alcuni dati recenti sull'avvicinamento di ChatGPT4 al livello di una AGI sono in [Bubeck *et al.*, 2023].
[12] Un buon testo di riferimento sul tema è [Jurafsky - Martin, 2023].

il *large language model* (LLM) vero e proprio: un modello di correlazioni statistico-probabilistiche fra *token*, ciascuno dei quali è rappresentato attraverso un'ampia matrice di valori numerici. In tal modo a ogni *token* viene associato ('vettorializzazione') uno spazio astratto e multidimensionale che esprime, in maniera puramente numerica, i contesti d'uso e le relazioni del *token* nel corpus: *token* con 'usi' simili, e dunque presumibilmente con significati vicini, corrisponderanno a vettori che avranno, almeno per alcune delle dimensioni, valori numerici abbastanza vicini; lo stesso avverrà, rispetto ad altre dimensioni, per *token* frequentemente usati insieme.

La costruzione dei vettori per ogni *token* – che, come si è detto, avviene nella fase di addestramento del modello – fornisce quello che è chiamato *embedding*: una rappresentazione che in sostanza cerca di coglierne, trasformandole in valori numerici, le modalità d'uso nel linguaggio. Va notato che le 'dimensioni' del vettore – che possono essere anche centinaia – sono puramente astratte e non corrispondono necessariamente (anzi, di regola non corrispondono affatto) alle categorie grammaticali o semantiche che utilizzeremmo noi per classificare una parola o un morfema.

Il modello così costruito sarà poi utilizzato per generare, a partire dal *prompt* dell'utente, la risposta del sistema, con un meccanismo detto *sequence-to-sequence*: partendo da una sequenza di simboli in ingresso viene generata una sequenza di simboli in uscita. Ma come funziona questo processo?

Inizialmente, per compiti simili erano utilizzate soprattutto le cosiddette 'reti neurali ricorrenti' (RNN)[13]. A differenza di una rete fatta di più strati di percettroni (*multi-layered perceptrons* o MLP[14]), in cui l'informazione viene elaborata con un movimento sempre 'in avanti' da uno strato all'altro (per questo si parla anche di *feedforward neural network*), nelle RNN è possibile prevedere cicli di rielaborazione dell'informazione da parte dello stesso strato della rete: questo permette – fra l'altro – di 'ridurre l'errore' in maniera molto più efficace. Tuttavia, nelle RNN l'analisi dei testi forniti come input (tanto a livello di corpus quanto a livello di *prompt*) e la produzione dell'output sono comunque fatti una parola alla volta. Reti di questo tipo hanno problemi di 'memoria semantica' (soprattutto nei contesti più lunghi, la pura associazione statistica di parole fornita attraverso l'*embedding* non basta a conservare la coerenza semantica del testo prodotto) e problemi di costi computazionali (il lavoro puramente sequenziale sfrutta male l'uso parallelo di più processori, indispensabile per lavorare su corpora assai ampi e su reti neurali molto complesse).

Il successivo (e fondamentale) passo in avanti sulla strada verso i sistemi generativi odierni viene fatto nel 2017, con la pubblicazione da parte di un gruppo di ricercatori impegnati nei laboratori di intelligenza artificiale di Google di un articolo che è negli ultimi anni probabilmente il più citato del settore: *Attention is all you need* [Vaswani *et al.* 2017]. È questo articolo che introduce una architettura di rete molto più efficace delle RNN: quella basata su *transformer*[15]. In questo caso, i *token* non sono più esaminati solo sequenzialmente ma anche tenendo conto del loro contesto, attraverso un meccanismo di 'attenzione' che 'pesa' i valori dei vettori di ogni *token* in funzione dei valori di ciascuno degli altri *token* del contesto. Nel farlo lavora, partendo dal vettore che rappresenta l'*embedding* del *token*, anche con tre vettori aggiuntivi, denominati – per analogia con le tecniche di ricerca in un database – *query*, *key* e *value*: non entreremo qui nel dettaglio del loro funzionamento, legato al processo di iterazione di cui parleremo fra un attimo.

Oltre a facilitare la disambiguazione di parole polisemiche (se il contesto di occorrenza di una parola come 'pesca' contiene anche termini come 'frutto' o 'succo', questo produrrà – partendo dal vettore iniziale – dei vettori pesati con valori più vicini a quelli di parole come 'arancia' o 'mela'; se invece il contesto contiene i termini 'pesce' o 'rete', questo produrrà dei vettori pesati con valori più vicini a quelli di parole come 'caccia' o 'sport'), questo metodo funziona molto meglio su input lunghi e produce sistemi con una 'memoria semantica' assai migliore. Inoltre, il meccanismo di attenzione può essere ripetuto più volte (*multi-head attention*) per dar conto, attraverso pesi diversi, di forme diverse di attenzione, alcune delle quali riguarderanno la semantica, altre la sintassi della frase: ad esempio, dopo un articolo ci si aspetta probabilmente – ma non necessariamente – un nome, e il sistema, in buona sostanza, dedicherà 'attenzione' anche a questi aspetti. Fermo restando che anche in questo caso la distinzione fra semantica e sintassi è solo un nostro modo possibile di guardare a quelle che per il sistema sono solo relazioni numeriche fra vettori, quasi mai direttamente interpretabili attraverso le nostre categorie linguistiche abituali.

Nel funzionamento del sistema, l'architettura basata sui *transformer* è applicata più volte, sequenzialmente, sia

[13] Per una introduzione al tema si veda [Glassner, 2021, capitolo 19].
[14] Per una introduzione al tema si veda il classico [Haykin, 1999, capitoli 3-4].
[15] Per una introduzione al tema si veda [Glassner, 2021, capitolo 20]. Una presentazione ancor più accessibile (e più dettagliata della sintesi necessariamente estrema che propongo in questa sede) può essere fornita da un video disponibile nel canale YouTube di Arkar Min Aung, del Worcester Polytechnic Institute, all'indirizzo <https://youtu.be/g2BRIuln4uc>.

nella codifica dell'input sia nella produzione della risposta. Di questo processo possono far parte due moduli diversi, denominati rispettivamente *encoder* e *decoder*. Non entrerò qui nel dettaglio del loro funzionamento: per avere un'idea del loro ruolo può però essere utile ricordare che nelle maggior parte delle intelligenze artificiali generative dedicate alla traduzione, l'*encoder* si occupa specificamente della rappresentazione attraverso vettori del testo ricevuto come input, e il *decoder* della generazione dell'output a partire dalla rappresentazione prodotta dall'*encoder*. La famiglia di modelli composta da T5 (*text to text transfer transformer*) e dai suoi successori usa una architettura di questo tipo, che unisce *encoder* e *decoder*. Ma è possibile avere anche *transformer* che si concentrano soprattutto sull'aspetto della rappresentazione e analisi del testo (e usano quindi solo *encoder*), o *transformer* che si occupano soprattutto della generazione di testo (e usano solo *decoder*). La scelta del modello più funzionale (*encoder-decoder*, solo *encoder* o solo *decoder*) dipenderà in parte dai nostri obiettivi.

Così, ad esempio, BERT (Bidirectional Encoder Representations from Transformers) – uno dei primi modelli basati su *transformer*, proposto nel 2018 da un gruppo di ingegneri di Google guidato da Jacob Devlin – aveva l'obiettivo di creare un LLM capace di 'predire' sia un *token* dato il suo contesto, sia la frase successiva di un dato contesto (NSP: *next sentence prediction*). In questo caso, l'attenzione è posta soprattutto sull'analisi del testo, e il *transformer* era quindi costituito solo da *encoder*. Questo tipo di architettura funziona bene anche nei casi in cui ci interessi ad esempio la cosiddetta *sentiment analysis* (l'identificazione delle connotazioni emozionali di un testo), o l'identificazione dei diversi 'agenti' in un dialogo, o ancora la costruzione di sommari o parafrasi del testo o la sua analisi semantica. Da BERT è nata tutta una famiglia di modelli *encoder-only*, con finalità e struttura in parte diverse.

OpenAI, la società che ha prodotto la famiglia di modelli GPT, ha invece lavorato soprattutto su sistemi composti solo da *decoder*, e finalizzati in primo luogo alla produzione di testo. I modelli GPT lavorano in questo modo, e la loro architettura – fatta solo di *decoder* – ha l'obiettivo finale di predire il *token* successivo partendo dai *token* precedenti. Le parole generate man mano dai *decoder* diventano esse stesse parte dell'input usato per la produzione della parola successiva, il che consente di generare risposte che mantengono coerenza sintattica e semantica anche se sono molto più lunghe del *prompt* di partenza.

È interessante notare che anziché scegliere sempre, volta per volta, la parola che il sistema seleziona come 'più rilevante', la generazione del testo utilizza una ulteriore componente stocastica, selezionando a volte parole con punteggi leggermente più bassi. La frequenza di questi 'scarti' è chiamata 'temperatura' del sistema: si è visto che un sistema con temperatura pari a circa 0,8 fornisce risposte più interessanti di un sistema con temperatura 0, che seleziona sempre la parola ottimale [Wolfram, 2023]. L'IA generativa presente in Bing, il motore di ricerca di casa Microsoft, permette così di selezionare fra tre diverse temperature delle risposte, identificate non in forma numerica ma in forma più colloquiale e comprensibile per l'utente: risposte 'creative', 'equilibrate' e 'precise', che corrispondono a temperature via via più basse.

Può essere utile a questo punto un'osservazione incidentale: si dice (e si legge) spesso che i sistemi di intelligenza artificiale generativa di ambito linguistico, come GPT e ChatGPT, non usano semantica ma solo sintassi e statistica. Il concetto di semantica è in parte ambiguo, perché ne esistono (almeno) due interpretazioni abbastanza diverse: studio delle relazioni fra segni e mondo esterno (nella tradizione di [Morris, 1938]), e studio del significato (nella tradizione di [Bréal, 1897]). Se si adotta l'idea di Morris (che porta a considerare la sintassi come teoria generale delle relazioni fra segni), l'*embedding* può in effetti essere ricondotto all'ambito della sintassi: non guarda al mondo esterno, ma solo alle relazioni interne alla sfera della concreta produzione linguistica. Ma in genere chi considera solo sintattico il lavoro di questi sistemi sembra fare un'assunzione molto più forte e sembra ritenere che esso non abbia nulla a che fare con i significati. Se riflettete sulla sintesi del loro funzionamento proposta fin qui – pur se breve e lacunosa – vi accorgerete che questa assunzione è fuorviante: la considerazione dei contesti d'uso dei *token* nel corpus e il meccanismo di attenzione producono infatti – almeno per chi considera la sfera della semantica come legata alla sfera dei significati – una sorta di 'semantica quantitativa': si lavora certo sempre con numeri, ma questi numeri 'incorporano' un'enorme quantità di informazioni che noi considereremmo semantiche, assieme a un'enorme quantità di informazioni che noi considereremmo sintattiche, e di informazioni che probabilmente non sapremmo bene come classificare o interpretare. Se ricordiamo il forte collegamento fra significato e uso delle parole stabilito – a partire da Wittgenstein – dalla filosofia del linguaggio novecentesca, potremmo dire che i *transformer*, lungi dall'essere solo 'macchine statistiche', sono (anche) una sorta di formalizzazione dell'idea di significato come uso.

Ma torniamo al nostro sistema di intelligenza artificiale generativa. Al termine della fase di apprendimento non supervisionato, il modello può essere ulteriormente perfezionato: sia integrandolo con corpora più ristretti e

specifici (qualora si voglia lavorare su ambiti particolari), sia attraverso fasi di *supervised learning*, in cui coppie di input-output – frutto delle interazioni fra esseri umani e sistema – sono sottoposte al vaglio di addestratori umani che possono a loro volta approvarle (rafforzando in tal modo i pesi dei collegamenti attivati nel produrre l'output partendo dal particolare input considerato) o respingerle (indebolendo tali pesi). Anche se OpenAI non ha mai dichiarato esplicitamente in che modo le interazioni con gli utenti finali siano utilizzate nell'ulteriore addestramento del sistema, la possibilità di dare un giudizio sulla bontà o meno delle risposte fornite (pollice in su o pollice verso) ha evidentemente anche lo scopo di utilizzare il proprio bacino di utenza come livello ulteriore di *supervised learning*. Inoltre, per addestrare il sistema possono essere usati meccanismi di apprendimento in cui l'output è analizzato da un altro sistema di intelligenza artificiale (spesso attraverso il meccanismo delle cosiddette 'reti generative avversarie', o GAN, che cercano di discriminare fra output artificiale e output umano e che la rete originale deve cercare di ingannare), e in alcuni casi anche dallo stesso sistema (*self-supervised learning*). Infine, all'output possono essere (e di fatto vengono spesso) applicati dei filtri a valle di vario genere, ad esempio per riconoscere e inibire risposte considerate per vari motivi come potenzialmente non accettabili.

Tanto il *supervised learning* quanto i filtri a valle dovrebbero aiutare a limitare le cosiddette 'allucinazioni' delle intelligenze artificiali generative di questo tipo: la produzione di testi con informazioni erronee, o che propongono tesi socialmente o eticamente inaccettabili, o che sembrano manifestare emozioni o volontà autonoma del sistema. Le allucinazioni rappresentano naturalmente un problema particolarmente grave se e quando consideriamo un sistema di questo tipo anche come una fonte di informazioni o come uno strumento di mediazione informativa: un tema che ha particolare rilevanza in questa sede, e su cui vale la pena soffermarsi.

Allucinazioni, *bias* e problemi delle IA generative

Ho cercato di fornire un'idea del funzionamento di sistemi come GPT e ChatGPT, perché molto spesso le discussioni sulla loro natura, sul loro futuro e sul loro impatto sociale e culturale (prevedibilmente assai notevole) sembrano prescindere completamente dal loro effettivo funzionamento: quando va bene, ci si limita a spiegazioni completamente generiche, e non di rado si ha l'impressione che chi ne analizza i possibili effetti, benefici o nefasti, non sappia però bene di cosa stia parlando.

Abbiamo visto che, sostanzialmente, sistemi di questo tipo producono testi in forma predittiva. Si tratta, si è detto, di previsioni statistiche basate su grandi modelli linguistici e su un lungo addestramento, in parte autonomo e in parte supervisionato: non vi è dunque nessuna 'copiatura' meccanica delle informazioni incamerate attraverso il corpus di testi di partenza, e non vi è neanche un'operazione di estrazione dal corpus delle informazioni considerate più rilevanti rispetto al *prompt* dell'utente.

In altri termini, GPT o ChatGPT non funzionano come delle enciclopedie o come dei sofisticati motori di ricerca, ma come complessi oracoli statistici; proprio per questo, le loro 'allucinazioni' possono essere particolarmente insidiose. Ancora pochi mesi fa, se si chiedeva a ChatGPT 3.5 un articolo accademico, con bibliografia, sull'effetto dell'introduzione della luce elettrica nella Firenze medievale, il sistema elaborava quella richiesta non già ricercando informazioni rilevanti da fonti affidabili (che l'avrebbero portato auspicabilmente a rispondere che nella Firenze medievale non esisteva energia elettrica) ma 'costruendo' una risposta plausibile. Nel caso specifico, la risposta sosteneva che l'introduzione dell'energia elettrica a Firenze nel Medioevo aveva permesso fra l'altro un allungamento degli orari lavorativi, l'illuminazione notturna della città, e lo sviluppo di nuove forme di arti visuali. Accompagnando queste considerazioni con una bibliografia che comprendeva un saggio di Leonardo da Vinci dedicato all'impatto dell'energia elettrica sulle arti[16]. Affascinante, ma ovviamente falso: qualcosa di simile alla risposta che avrebbe potuto dare uno studente poco preparato alla domanda-trabocchetto di un docente, con il tentativo di produrre – in assenza di informazioni adeguate – una risposta comunque apparentemente plausibile.

ChatGPT 4 non fa più questo specifico errore: il sistema è più elaborato e complesso, utilizza una base testuale più ampia[17], ed è capace di 'costruire' (il meccanismo di produzione della risposta è comunque fondamentalmente lo stesso) la risposta giusta: «Mi dispiace, ma c'è un errore nella tua richiesta. L'elettricità non è stata introdotta durante il Medioevo a Firenze o in altre parti del mondo. L'elettricità come fonte di energia per l'illuminazione e l'alimentazione delle macchine è stata introdotta solo nel diciannovesimo secolo, molti secoli

[16] Le immagini relative a questa risposta – fornita da ChatGPT 3.5 il 20 dicembre 2022 – sono incluse nell'espansione multimediale di questo articolo.

[17] Una descrizione abbastanza dettagliata del sistema, con esempi di grande interesse, è in [OpenAI, 2023].

dopo il periodo che hai menzionato»[18]. È probabile, peraltro, che la capacità di correggere questo tipo di errori non sia solo il risultato del corpus di partenza più ampio e del maggior numero di parametri utilizzati dalla nuova versione, ma anche (o soprattutto) di addestramento supervisionato, legato non già a questo *prompt* particolare, ma all'esigenza di limitare gli effetti dei frequenti tentativi di ingannare il sistema da parte degli utenti. In altri termini, il sistema ha dovuto progressivamente 'imparare' a essere meno ingenuo nella costruzione delle risposte, aumentando ulteriormente il peso delle associazioni che si rivelano essere fattualmente corrette e indebolendo quello delle associazioni puramente inventate. La 'spinta' data in questo senso dall'addestramento supervisionato rinforza la capacità di fornire risposte del tipo "Mi spiace, ma c'è un errore nella tua richiesta" e di farle seguire da informazioni fattualmente corrette.
Un altro aspetto importante da considerare è la capacità di 'allargare' il contesto esaminato dal sistema nella generazione della risposta. Si è detto che la creazione dei LLM richiede una fase di addestramento lunga, condotta su molti processori paralleli e dunque ad alto consumo energetico: per questo la conoscenza interna di un modello è limitata al periodo in cui il corpus su cui è stato addestrato è stato costruito e sottoposto al sistema. Sappiamo così, ad esempio, che GPT 3 non aveva accesso a informazioni successive a fine 2021. Tuttavia, è possibile permettere al sistema di utilizzare informazioni più recenti addestrandolo a lavorare anche su un contesto aggiunto in un secondo momento, o ricavato in tempo reale dalla rete attraverso ricerche condotte sulla base del *prompt* ricevuto. Le informazioni 'nuove' non entreranno nella costruzione iniziale dei pesi del modello, ma potranno essere gestite come una sorta di 'estensione' (o, appunto, contesto) del *prompt* dell'utente, entrando dunque nella produzione delle risposte. È quanto ha fatto ad esempio il motore di ricerca di casa Microsoft, Bing. Se questo allargamento del contesto avviene sfruttando fonti verificate e validate, la produzione di risposte corrette sarà più facile.
Queste strategie non possono eliminare completamente le 'allucinazioni' – che, come abbiamo visto, sono in parte il prodotto del meccanismo di costruzione delle risposte utilizzato dalle IA generative – ma possono sicuramente limitarle. È importante tener presente, tuttavia, che le allucinazioni non sono l'unico problema che possono presentare le risposte fornite da sistemi di IA generativa. In particolare, è stato già verificato[19] che questi sistemi possono riprodurre nelle loro risposte *bias* sistematici legati al corpus su cui sono stati addestrati[20] (ad esempio, se chiediamo a ChatGPT di costruire una storia, è molto probabile che utilizzi nomi occidentali, e dia per scontato che la protagonista o il protagonista non appartenga a minoranze etniche o religiose; se chiediamo una storia su un'attrice, probabilmente sarà descritta come bellissima). Peraltro, l'esatta natura e composizione di questi corpora – che inizialmente erano per lo più pubblicamente disponibili[21] – è diventata man mano meno trasparente, rendendo più difficile identificare i *bias* che essi possono presentare. E naturalmente dei *bias* possono essere introdotti anche nella fase di apprendimento supervisionato, che può risentire dei pregiudizi di chi interpreta e valuta gli output. Paradossalmente, la stessa tendenza a cercare di favorire la produzione di risposte 'politicamente corrette' e inclusive può essere percepita da alcuni come *bias*, come mostra l'acceso dibattito nato negli Stati Uniti sull'orientamento 'democratico' di ChatGPT e sulla sua avversione per le politiche repubblicane[22]. Ed è sicuramente preoccupante la possibilità di costruire LLM programmaticamente orientati dal punto di vista politico o ideologico, o addestrati su testi e da parte di istruttori che desiderano utilizzarli per difendere tesi false o inaccettabili. In linea di principio la costruzione di LLM volutamente 'cattivi' [OpenAI, 2023] è non solo possibile ma anche relativamente facile, almeno finché ci si limita a modelli non troppo costosi dal punto di vista computazionale (e sappiamo che i costi computazionali tendono a diminuire nel tempo).
Ma quello dei *bias* è solo uno dei molti aspetti proble-

[18] ChatGPT 4, risposta fornita all'autore l'8 aprile 2023.
[19] Si veda ad esempio [Silva - Tambwekar - Gombolay, 2021].
[20] Per *bias* si intende una distorsione sistematica nei dati o nei modelli utilizzati, che porta il sistema a produrre risultati a loro volta distorti o inaffidabili. Per alcuni esempi – fra i molti possibili – si veda [Bender *et al.*, 2021].
[21] È il caso ad esempio di BookCorpus, su cui sono state addestrate le prime versioni sia di GPT sia di BERT (un corpus di circa 11.000 libri di vario genere diffusi gratuitamente sul sito Smashwords e scritti da autori non legati a case editrici commerciali), realizzato nel 2015 da ricercatori dell'Università di Toronto e del MIT e reso disponibile fino a pochi anni fa dal sito web dell'Università di Toronto (è ancora possibile reperirlo in rete, ma non più in forma ufficiale), o di CommonCrawl Corpus, il corpus aperto e gratuito realizzato dall'organizzazione no-profit CommonCrawl (<https://commoncrawl.org>), che raccoglie il testo di milioni di pagine web e una cui versione filtrata (come è facile immaginare, il corpus CommonCrawl comprende anche contenuti pornografici o inaffidabili) è stata usata nell'addestramento di GPT3. Un'interessante ricostruzione della storia di BookCorpus è in [Bandy - Vincent, 2021]. È chiaro che entrambi questi corpora, tutt'altro che 'asettici', possono presentare e di fatto presentano *bias* sistematici [Birhane - Prabhu - Kahembwe, 2021].
[22] Per un esempio si veda [Mitchell, 2023]. Per un curioso parallelo relativo alla politica italiana si veda [Signorelli, 2023].

matici su cui, pur senza demonizzare la ricerca nel campo delle IA generative e i suoi affascinanti risultati, è indispensabile riflettere. Sono così preoccupanti, per fare solo qualche ovvio esempio, la progressiva 'chiusura' sia dei corpora usati nell'addestramento sia delle metodologie, dei modelli e delle architetture usate[23]; le politiche (o l'assenza di politiche) di gestione dei dati personali forniti dagli utenti attraverso i loro *prompt*; la questione del copyright di contenuti generati sulla base di un addestramento su dati creati da persone che non sono necessariamente consapevoli dell'inclusione di testi o immagini creati da loro nel corpus utilizzato, e che non partecipano né alla definizione delle relative politiche né agli utili derivanti dal loro sfruttamento economico; il rischio che la concorrenza sregolata fra società che producono sistemi di IA generativa si traduca in una eccessiva accelerazione nel loro sviluppo senza il tempo necessario alla riflessione e all'elaborazione di adeguate garanzie etiche, così come di valutazioni del loro impatto non solo culturale ma anche politico, sociale, economico e occupazionale[24]. Si tratta di problemi di enorme rilievo, che considerato il focus di questo intervento posso qui solo ricordare; alcuni di essi sono discussi più approfonditamente in altri interventi in questo fascicolo, su molti altri il dibattito è già avviato, ed è bene proceda con tutto l'approfondimento necessario.

Conclusioni

Ho già osservato che le IA generative 'costruiscono' risposte anziché limitarsi a selezionare e riproporre contenuti già esistenti in rete. Questo ha vantaggi e svantaggi: abbiamo appena ricordato alcuni dei potenziali svantaggi e problemi, ma va detto che la loro capacità di produzione di testo semanticamente e contestualmente adeguato è realmente impressionante, e ci sono pochi dubbi sul fatto che sistemi come quelli qui considerati saranno in grado abbastanza presto di superare il test di Turing (probabilmente, possono già ingannare un esaminatore privo di competenze specifiche di alto livello).
Finché le allucinazioni resteranno un problema frequente, difficilmente vorremmo affidare a sistemi di questo tipo funzioni complesse di mediazione informativa, o la capacità di far seguire alla produzione linguistica anche azioni autonome capaci di avere effetti sul mondo reale. Così – in questo momento – GPT, ChatGPT o i sistemi in parte analoghi in sviluppo presso altre aziende (come Bard, il sistema su cui lavora Google a partire da un LLM chiamato LaMDA, o il LLM di Meta, chiamato LLaMA) non sarebbero adatti a svolgere funzioni come quelle del reference bibliotecario (cosa pensereste di un bibliotecario che suggerisce come bibliografia articoli inesistenti?), difficilmente potrebbero produrre ricerca innovativa in ambito umanistico (mentre in ambito scientifico alcune IA sono già utilizzate in ambiti specifici, come la genomica, lo sviluppo di nuove proteine o come supporto alla programmazione di software), e non affideremmo loro il nostro numero di carta di credito. Ma non è affatto detto che queste limitazioni dureranno nel tempo: chi ritiene che alle IA generative manchi la 'creatività' umana tende forse a sopravvalutare la dose di creatività richiesta da molti compiti anche altamente astratti, comprese molte forme di ricerca accademica o di mediazione informativa. E probabilmente avrebbe difficoltà a descrivere il concetto di creatività in forme tali da portare a escludere in linea di principio che intelligenze artificiali generative possano replicarne i risultati.
Un discorso analogo si può fare per quanto riguarda la semantica. Ho già notato come uno sguardo appena più ravvicinato a questi sistemi porti a smentire l'idea diffusa di una macchina che lavori prescindendo dal piano dei significati: certo GPT o ChatGPT non 'capiscono' i testi nello stesso senso in cui li capiamo noi, non hanno coscienza o autocoscienza (ammesso e non concesso che sia chiaro cosa sia e come funzioni la nostra autocoscienza), non hanno intenzionalità (anche qui, ammesso e non concesso che sia chiara la natura dell'intenzionalità umana), e sembra difficile – almeno per ora – attribuire loro un ruolo di agenti autonomi [Lana, 2022]. Ma si tratta di motori semantici, non solo sintattici, e la loro capacità di produrre significati è radicata nella costruzione di modelli linguistici che non hanno solo un interesse pratico, ma anche un notevole interesse teorico. Modelli che in linea di principio potrebbero aiutarci a capire anche alcuni dei meccanismi di funzionamento della nostra produzione linguistica, peraltro in molti casi ancora abbastanza oscuri. Siamo così sicuri che modelli di generazione statistica non abbiano un ruolo anche nella 'nostra' competenza linguistica[25]? E siamo così sicuri che la nostra coscienza e la nostra autocoscienza

[23] Così, ad esempio, OpenAI, nata – come suggerisce il nome – con una politica non commerciale si è ormai trasformata in una azienda assai poco 'aperta'; più fedele alle politiche 'open' è per il momento LLaMA [Wolfe, 2023].
[24] Su questi ultimi aspetti è abbastanza impressionante la lettura di report come quello prodotto nel marzo 2023 da Goldman Sachs [Hatzius *et al.*, 2023].
[25] Può essere il caso di ricordare che i migliori algoritmi di attribuzione dei testi (quelli utilizzati quando dobbiamo cercare di attribuire un testo anonimo o 'sospetto' a un autore o all'altro) non utilizzano le componenti semanticamente e stilisticamente significati-

non siano a loro volta almeno in parte collegate a un sostrato di questo tipo? O, ancora, che l'intelligenza richieda necessariamente la coscienza, o una forma di coscienza analoga alla nostra?

Il problema del rapporto fra cervello e coscienza (su cui non intendo certo soffermarmi in questa sede: per una interessante introduzione recente si veda [Seth, 2022]) è una questione – un tempo solo filosofica, oggi anche neurofisiologica – aperta da secoli ma ben lontana dall'essere risolta: se non accettiamo forme di dualismo o spiritualismo, sembra assai difficile non considerare la coscienza come un fenomeno emergente a partire da un sostrato biologico. Ma come funzioni concretamente questo sostrato, come sia capace di processare e di produrre informazioni, come intrecci dati sensoriali, dati linguistici, significati, emozioni, intenzioni, è cosa che, nonostante gli indubbi progressi fatti dalle neuroscienze negli ultimi decenni, continua a essere abbastanza oscuro.

Sappiamo che i sistemi di intelligenza artificiale generativa sono abbastanza diversi da noi – non solo nel loro funzionamento, ma anche nell'assenza di connessione diretta con un corpo biologico – da suggerire che, almeno al momento, attribuire loro coscienza, o intenzioni, o una intelligenza analoga alla nostra, rappresenti una antropomorfizzazione non solo ingiustificata ma anche assai fuorviante. Sappiamo però anche che la loro capacità di produrre contenuti linguistici sintatticamente e semanticamente appropriati – una capacità tradizionalmente considerata appannaggio dell'intelligenza umana – li trasforma in agenti linguistici almeno parzialmente autonomi[26]. E sappiamo che almeno alcuni dei loro meccanismi di funzionamento interno sono per un verso, come abbiamo visto, opachi per gli stessi programmatori, e per altro verso forse non così radicalmente diversi da quelli che utilizzano anche alcune componenti di basso livello della nostra intelligenza: tanto da non permettere di escludere del tutto, anche per il futuro, prospettive per ora relegate al campo della fantascienza.

RIFERIMENTI BIBLIOGRAFICI

Abraham, 2002 = Abraham Tara H., *(Physio)logical circuits: the intellectual origins of the McCulloch-Pitts neural networks*, «Journal of the history of the behavioral sciences», 38 (2002), n. 1, p. 3-25, DOI: 10.1002/jhbs.1094.

Bandy - Vincent, 2021 = Bandy Jonh - Vincent Nicholas, *Addressing "documentation debt" in machine learning: a retrospective datasheet for BookCorpus*, in *Proceedings of the Neural Information Processing Systems, Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks 2021)*, edited by Joaquin Vanschoren, Serena Yeung, La Jolla (CA), Neural Information Processing Systems, 2021, <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/54229abfcfa5649e7003b83dd4755294-Abstract-round1.html>.

Bender *et al.*, 2021 = Bender Emily M. [*et al.*], *On the dangers of stochastic parrots: can language models be too big?,* in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, New York, Association for Computing Machinery, 2021, p. 610-623, DOI: 10.1145/3442188.3445922.

Birhane - Prabhu - Kahembwe, 2021 = Birhane Abeba - Prabhu Vinay Uday - Kahembwe Emmanuel, *Multimodal datasets: misogyny, pornography, and malignant stereotypes*, «arXiv», 5 October 2021, art. arXiv:2110.01963, DOI: 10.48550/arXiv.2110.01963.

Bostrom, 2014 = Bostrom Nick, *Superintelligence: paths, dangers, strategies*, Oxford, Oxford University Press, 2014.

Bréal, 1897 = Bréal Michel, *Essai de sémantique*, Paris, Hachette, 1897 (trad. it. *Saggio di semantica*, Napoli, Liguori 1990).

Bubeck *et al.*, 2023 = Bubeck Sébastien [*et al.*], *Sparks of artificial general intelligence: early experiments with GPT-4*, «arXiv», 13 April 2023, art. arXiv:2303.12712v5, DOI: 10.48550/arXiv.2303.12712.

Cave - Dihal - Dillon, 2020 = *AI narratives: a history of imaginative thinking about intelligent machines*, edited by Stephen Cave, Kanta Dihal, Sarah Dillon, Oxford, Oxford University Press 2020.

Chandra, 2018 = Chandra Akshay L., *McCulloch-Pitts neuron - mankind's first mathematical model of a biological neuron*, «Medium», 24 July 2018, <https://towardsdatascience.com/ mcculloch-pitts-model-5f-df65ac5dd1>.

Floridi, 2023 = Floridi Luciano, *AI as agency without in-*

ve della frase che potremmo ragionevolmente aspettarci come criteri distintivi per la produzione linguistica umana (come lessico, lunghezza media delle frasi, uso della punteggiatura ecc.), ma un dato puramente sintattico: la frequenza di *n*-grammi, e in particolare di bigrammi e trigrammi, risultato della scomposizione del testo in gruppi di *n* caratteri (una metodologia presentata per la prima volta in forma sistematica in [Peng *et al.*, 2003]). Peraltro, i programmi di attribuzione dei testi rappresentano un caso particolare di programmi di classificazione (l'attribuzione del testo a questo o quell'autore all'interno di un insieme di autori possibili): si tratta dunque di un compito per il quale possono essere utilizzati efficacemente anche sistemi basati su reti neurali.

[26] Sul rapporto fra 'agency' e intelligenza in sistemi come ChatGPT si veda [Floridi, 2023].

telligence: on ChatGPT, large language models, and other generative models, «Philosophy & technology», 36 (2023), n. 1, p. 1-7, DOI: 10.1007/s13347-023-00621-y.

Glassner, 2021 = Glassner Andrew, *Deep learning: a visual approach*, San Francisco, No Starch Press, 2021.

Hatzius *et al.*, 2023 = Hatzius Jan [*et al.*], *The potentially large effects of artificial intelligence on economic growth*, «Goldman Sachs Economic Research», 26 March 2023.

Haykin, 1999 = Haykin Simon, *Neural networks: a comprehensive foundation*, 2. ed., Upper Saddle River (NJ), Prentice Hall, 1999.

Hoffmann, 2022a = Hoffmann Christian H., *A philosophical view on singularity and strong AI*, «AI & society», 2022, DOI: 10.1007/s00146-021-01327-5.

Hoffmann, 2022b = Hoffmann Christian H., *The quest for a universal theory of intelligence: the mind, the machine, and singularity hypotheses*, Berlin, De Gruyter, 2022.

Jurafsky - Martin, 2023 = Jurafsky Daniel - Martin James H., *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*, 3. ed. draft, 7 January 2023, <https://web.stanford.edu/~jurafsky/slp3> (la seconda edizione del volume è stata pubblicata da Prentice Hall nel 2009).

Kurzweil, 2005 = Kurzweil Ray, *The singularity is near: when humans transcend biology*, London, Viking, 2005.

McCarthy *et al.*, 1955 = McCarthy [*et al.*], *A proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955*, «AI magazine», 27 (2006), n. 4, p. 12-14, DOI: 10.1609/aimag.v27i4.1904.

McCulloch - Pitts, 1943 = McCulloch Warren S. - Pits Walter, *A logical calculus of the ideas immanent in nervous activity*, «The bulletin of mathematical biophysics», 5 (1943), n. 4, p. 115-133, DOI: 10.1007/BF02478259.

Minsky - Papert, 1969 = Minsky Marvin - Papert Seymour, *Perceptrons: an introduction to computational geometry*, Cambridge (MA), MIT Press, 1969.

Mitchell, 2023 = Mitchell Alex, *'Wild West' ChatGPT has 'fundamental flaw' with left bias*, «New York Post», 15 February 2023, <https://nypost.com/2023/02/15/wild-west-chatgpt-has-fundamental-flaw-with-left-bias>.

Morris, 1938 = Morris Charles, *Foundations of the Theory of signs*, Chicago (IL), The University of Chicago Press, 1938 (trad. it. *Fondamenti di una teoria dei segni*, Torino, Paravia, 1955).

Norvig - Russell, 2021 = Russell Stuart - Norvig Peter, *Artificial intelligence: a modern approach*, 4. ed., Harlow, Pearson, 2021.

Numerico, 2021 = Numerico Teresa, *Big data e algoritmi: prospettive critiche*, Roma, Carocci, 2021.

OpenAI, 2023 = OpenAI, *GPT-4 technical report*, 27 March 2023, <https://cdn.openai.com/papers/gpt-4.pdf>.

Peng *et al.*, 2003 = Peng Fuchun [*et al.*], *Language independent authorship attribution with character level n-grams*, in *10th Conference of the European Chapter of the Association for Computational Linguistics*, *Budapest, April 2023*, edited by Ann Copestake, Jan Hajič, Stroudsburg (PA), Association for Computational Linguistics, 2003, p. 267-274.

Roncaglia, 2014 = Roncaglia Gino, *Computer che copiano: test di Turing, web corpora e filtraggio collaborativo*, in *Per il centenario di Alan Turing fondatore dell'informatica: Roma, 22 novembre 2012: convegno*, Roma, Scienze e lettere, 2014, p. 189-201, <http://hdl.handle.net/2067/2614>.

Searle, 1980 = Searle John R., *Minds, brains, and programs*, «Behavioral and brain sciences», 3 (1980), n. 3, p. 417-424 (trad. it. *La mente è un programma?*, «Le scienze», 259 (1990), <https://www.lescienze.it/archivio/articoli/1990/03/01/news/la_mente_e_un_programma_-545024>).

Seth, 2022 = Seth Anil, *Being you: a new science of consciousness*, London, Faber & Faber, 2022 (trad. it. *Come il cervello crea la nostra coscienza*, Milano, Raffaello Cortina, 2023).

Signorelli, 2023 = Signorelli Andrea Daniele, *ChatGPT vota PD?*, «Wired», 25 febbraio 2023, <https://www.wired.it/article/chatgpt-pd-idee-politiche-destra-sinistra-intelligenza-artificiale>.

Silva - Tambwekar - Gombolay, 2021 = Silva Andrew - Tambwekar Pradyumna - Gombolay Matthew, *Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers*, in *The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by Kristina Toutanova [*et al.*], Stroudsburg (PA), Association for Computational Linguistics, 2021, p. 2383-2389, DOI: 10.18653/v1/2021.naacl-main.189.

Turing, 1950 = Turing Alan M., *Computing machinery and intelligence*, «Mind», 59 (1950), n. 236, p. 433-460, <https://redirect.cs.umbc.edu/courses/471/papers/turing.pdf>.

Vaswani *et al.*, 2017 = Vaswani Ashish [*et al.*], *Attention is all you need*, in *Advances in Neural Information Processing Systems 30: 31st Annual Conference on Neural Information Processing Systems (NIPS 2017)*: *Long Beach, California, USA, 4-9 December 2017,* edited by Ulrike von Luxburg [*et al.*], La Jolla (CA), Neural Information Proessing Systems, 2017, p. 5998-6008, <https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.

Wolfe, 2023 = Wolfe Cameron R., *LLaMA*: *LLMs for everyone! High-performing language models that are fully open-source*, «Deep (learning) focus», 10 April 2023, <https://cameronrwolfe.substack.com/p/llama-l-lms-for-everyone>.

Wolfram, 2023 = Wolfram Stephen, *What is ChatGPT doing … and why does it work?*, «Stephen Wolfram Writings», 14 February 2023, <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/>.

ABSTRACT

Le intelligenze artificiali generative sono da diversi mesi al centro di una notevole attenzione mediatica, ed è abbastanza diffusa la previsione che il loro sviluppo porterà rapidamente a cambiamenti anche radicali in molti ambiti professionali, incluso il mondo della mediazione informativa.

Si tratta di una previsione giustificata? E quali sono gli sviluppi che possiamo aspettarci al riguardo? L'articolo intende presentare sinteticamente il contesto all'interno del quale si è sviluppato il lavoro su questi sistemi, i meccanismi di funzionamento e le caratteristiche di alcuni di essi (in particolare di quelli basati sulla generazione di testi attraverso transformer, come GPT e ChatGPT), i principali problemi riscontrati e una prima, assai parziale, riflessione sull'impatto che potranno avere nel futuro

GENERATIVE ARTIFICIAL INTELLIGENCE AND INFORMATION MEDIATION: AN INTRODUCTION

Generative artificial intelligence has been receiving considerable media attention for several months, and there is a widespread prediction that its development will rapidly bring about even radical changes in many professional fields, including the world of information mediation.

Is this prediction justified? And what are the developments that we can expect in this regard? This paper aims to briefly present the context in which work on these systems has developed, their mechanisms of operation, and the characteristics of some of them (in particular, those based on text generation through transformers, such as GPT and ChatGPT), the main problems encountered, and an initial, albeit partial, reflection on the impact they may have in the future.

Visita la piattaforma per scoprire i contenuti aggiuntivi

http://bibliotecheoggitrends.it