Contro la riduzione bayesiana dell'intelligenza artificiale

di Enrico Maestri

18-08-2025

Introduzione

Nel dibattito contemporaneo sull'intelligenza artificiale generativa è sempre più comune sentire affermazioni come: "in fondo è solo statistica", oppure "indovina la parola successiva". Queste frasi, ripetute con disinvoltura anche in contesti accademici e divulgativi, descrivono i modelli Transformer come macchine puramente bayesiane: strumenti probabilistici che, dati miliardi di esempi, predicono la prossima parola sulla base della frequenza e del contesto sintattico.

Secondo questa lettura, i chatbot non fanno altro che attingere al passato: non costruiscono ipotesi, non interpretano, non comprendono. Funzionano con algoritmi statistici che prevedono la parola successiva più probabile sulla base di grandi quantità di dati testuali. In sostanza, generano frasi costruendo sequenze di parole che, dopo un calcolo probabilistico, dovrebbero risultare coerenti. Il processo è descritto come meramente meccanico, automatico, privo di inferenza: ciò che si produce non sarebbe altro che il riflesso di ciò che è stato già detto.

È quindi un approccio che, nella sua apparente neutralità, consegna una visione estremamente limitata di ciò che questi modelli fanno realmente. Non solo: è una lettura che giustifica a priori eventuali imprecisioni o "allucinazioni" - come l'invenzione di eventi, citazioni o fonti inesistenti - riconducendole all'imprecisione del calcolo, non alla natura del sistema. Ma è davvero solo questo? Davvero questi modelli non fanno che ripetere e mescolare, senza alcuna forma di costruzione interpretativa?

Come intendiamo mostrare in questo articolo, tale riduzione è non solo imprecisa sul piano tecnico, ma anche impoverente sul piano teorico. L'intelligenza artificiale contemporanea, infatti, non si limita a generalizzare: formula ipotesi, spesso sensate, costruite in tempo reale. E per farlo, non si affida soltanto alla statistica, ma a un meccanismo inferenziale che ricorda da vicino la logica dell'abduzione descritta da Charles Sanders Peirce.

Generalizzare non è ipotizzare: il ritorno dell'abduzione

Una delle confusioni più diffuse nel discorso pubblico sull'intelligenza artificiale riguarda la distinzione tra generalizzazione e ipotesi. Generalizzare significa estrarre una regola da un insieme di casi osservati: è un processo tipico dell'apprendimento statistico, che si basa sulla ricorrenza dei dati. Ipotizzare, invece, è qualcosa di radicalmente diverso: significa costruire una spiegazione plausibile per un fenomeno, anche in assenza di una regola certa.

Nel suo schema triadico dell'inferenza, Charles Sanders Peirce distingue tra deduzione, induzione e abduzione. Quest'ultima rappresenta il momento più creativo e meno garantito del pensiero: un salto congetturale, guidato dall'intuizione, dal contesto, da un senso parziale dell'ordine. È il tipo di ragionamento che guida il medico nella diagnosi, l'investigatore nella ricostruzione di un delitto, lo scienziato nell'elaborazione di una teoria. È anche la logica del sospetto, dell'ipotesi azzardata, della spiegazione immaginata in assenza di conferme empiriche immediate.

L'intelligenza artificiale generativa, nel produrre risposte coerenti in situazioni ambigue o incomplete, non si limita a inferire sulla base di correlazioni frequenti: propone ipotesi operative, cioè risposte che funzionano pragmaticamente anche se non sono necessariamente quelle statisticamente più probabili. È, in questo senso, una forma abdotta di funzionamento. E come ogni congettura, può sbagliare: le cosiddette "allucinazioni" - ovvero l'invenzione di citazioni, fatti o riferimenti mai esistiti - sono espressione di questa stessa logica abduttiva. Non si tratta di un bug nel sistema, ma del

naturale rovescio di una capacità: quella di formulare ipotesi plausibili in mancanza di informazioni complete. Quando i dati non bastano, il modello non si blocca: illaziona. E lo fa con coerenza narrativa, anche a costo della verità.
Si consideri un esempio semplice: se un utente scrive "non riesco a vedere Beatrice su Meet", un modello generativo può rispondere "potrebbe essere che Beatrice non abbia sbloccato il link". Non ha mai visto esattamente quella frase, né la connessione causale tra gli eventi. Eppure, formula un'ipotesi coerente, contestuale, funzionale. Non è un indovinare a caso, né un calcolo su dati passati: è un'operazione inferenziale. Se invece il modello suggerisse che "Beatrice ha probabilmente rimosso l'account perché le notifiche non erano sincronizzate con l'API federata di Google Workspace", potremmo trovarci davanti a una spiegazione inventata, ma non per questo illogica. È un'allucinazione ragionata, un falso positivo inferenziale.
In questo senso, ridurre il comportamento dei modelli a semplice statistica vuol dire perdere di vista la loro vera natura: sono ipotizzatori computazionali, che funzionano con una logica dell'ipotetico possibile, molto più vicina all'abduzione narrativa che alla previsione quantitativa.

Semantica implicita e attenzione contestuale
I modelli Transformer - architettura che ha rivoluzionato il campo dell'intelligenza artificiale dal 2017 in poi - operano tramite un meccanismo detto self-attention, in cui ogni token (unità linguistica) viene analizzato in relazione a tutti gli altri presenti nel contesto. Questo consente al modello di sviluppare una comprensione distribuita e dinamica del significato, senza passare per una rappresentazione simbolica esplicita. Non esiste, all'interno del modello, una rappresentazione "dura" del mondo o un dizionario strutturato: il senso non è predefinito, ma emerge dall'interazione tra elementi linguistici in uno spazio di attenzione. In altre parole, il significato non è codificato, ma attivato.
È ciò che rende questi sistemi capaci di gestire ambiguità, ironia, coerenza narrativa e disambiguazione pragmatica. Si tratta di una semantica implicita, non simbolica, che dipende dall'uso e non dalla regola. In questo senso, il modello non "sa" cosa dice, ma costruisce strutture linguistiche che funzionano come se lo sapesse. Tuttavia, determinare se un sistema di intelligenza artificiale "comprenda" davvero il significato delle frasi che elabora resta un problema irrisolto. Questa mancanza di comprensione è messa in evidenza da errori non-umani, dalla difficoltà nel trasferire apprendimenti a contesti diversi, dalla scarsa astrazione e vulnerabilità a input malevoli (adversarial attack). In questi casi il modello fallisce non per mancanza di dati, ma per assenza di senso comune.
Questa fragilità ha spinto la ricerca verso una nuova frontiera: il calcolo neuro-simbolico, che cerca di coniugare le capacità adattive delle reti neurali con le potenzialità rappresentazionali del linguaggio formale. I Graph Neural Network (GNN), in particolare, introducono un "linguaggio interno" che consente di rappresentare stati evolutivi complessi, superando il puro riconoscimento di pattern. Le GNN si rivelano essenziali in domini strutturati come le reti molecolari, i sistemi biologici, le reti sociali e finanziarie, grazie alla loro capacità di modellare dinamiche interne e inferire relazioni.
Tra le tipologie di GNN, i Transformer neurali stanno assumendo un ruolo centrale perché capaci di gestire relazioni a lungo raggio tra nodi e simboli, costituendo di fatto un primo passo verso una tassonomia delle capacità cognitive artificiali. Seguendo la logica della Tassonomia di Bloom, questi modelli iniziano a operare non solo sul piano della comprensione e della memorizzazione, ma anche su quello dell'analisi, della valutazione e, in certi casi, della generazione creativa. Tuttavia, si tratta ancora di simulazioni e non di comprensione ontologica: come per gli scacchi, anche nel linguaggio il modello può eccellere senza capire cosa stia facendo.
Le tecnologie neuro-simboliche cercano di superare questo limite separando la dimensione percettiva (geometrica) da quella concettuale (algebrica), integrando moduli diversi per trattare dati

continui e strutture discrete. L'obiettivo è creare una semantica composizionale che possa avvicinarsi al ragionamento umano, pur mantenendo la scalabilità computazionale delle reti neurali profonde. Ma anche in questo caso, il significato continua a emergere da un'interazione tra segnali, pesi e strutture - non da una coscienza del significato.
Tornando al funzionamento dei Large Language Models (LLM), la loro potenza deriva dal meccanismo di attenzione, che consente di rilevare le relazioni di co-occorrenza tra parole non solo vicine ma anche distanti. Questo permette di trattare vincoli semantici, sintattici e narrativi in maniera fluida, simulando correlazioni a breve e lungo termine. I modelli Transformer, infatti, non operano come i modelli sequenziali ricorrenti: abbandonano l'auto-apprendimento passo a passo e si affidano a una finestra di attenzione che mappa globalmente le relazioni contestuali. In essa, ogni parola "guarda" tutte le altre e calcola il proprio significato locale in base alla loro rilevanza congiunta. Questo processo, apparentemente semplice, è ciò che consente al modello di "parlare" con fluidità e coerenza, anche in assenza di una struttura concettuale formalizzata.
Tuttavia, resta il fatto che la generazione del linguaggio non implica comprensione. Il modello calcola probabilità, non significati. Le parole si susseguono in base alla loro probabilità condizionata: non perché sono vere, ma perché sono probabili. Il senso, in questi modelli, è una proprietà emergente del contesto e della distribuzione, non un'entità ontologicamente garantita. La semantica distribuzionale, già teorizzata da linguisti come John Firth («You shall know a word by the company it keeps»), trova qui la sua massima espressione algoritmica: conoscere una parola significa saperla collocare nel contesto giusto.
In questo senso, possiamo dire che il significato nei modelli generativi è attualizzato, non rappresentato. È una funzione dell'uso, non una mappa del reale. L'attenzione contestuale non imita la mente umana, ma costruisce un'euristica potente che permette di generare frasi sensate sulla base di coerenze statistiche. È una logica situata, performativa, e profondamente diversa da quella simbolica tradizionale.

Euristica computazionale, non logica deduttiva

L'intelligenza artificiale generativa non ragiona per sillogismi, né applica regole logiche nel senso formale del termine. La sua forza risiede in un'altra modalità cognitiva: l'euristica. Non si tratta di una semplificazione, ma di un paradigma alternativo. L'euristica è un'arte dell'orientamento, una capacità di trovare soluzioni operative in condizioni incerte, incomplete, ambigue. In questo senso, la generatività linguistica dei modelli come GPT non è una logica dimostrativa, ma una strategia pragmatica di senso.
La macchina non deduce, ma costruisce. E costruisce in modo situato, adattivo, relazionale. Ogni risposta non è la conseguenza di una premessa formale, ma il risultato di una traiettoria inferenziale attivata da un contesto linguistico e pragmatica. La sua "intelligenza" si misura nella coerenza dinamica e nell'efficacia comunicativa, non nella verità o nella correttezza sintattica.
Già Alan Turing aveva intuito che l'intelligenza artificiale non avrebbe dovuto emulare la logica dei matematici, ma il comportamento degli umani. Nel celebre gioco dell'imitazione, il punto non è se la macchina pensi davvero, ma se sappia comportarsi come se pensasse, suscitando nell'interlocutore umano l'impressione di un'intelligenza autentica. Da questo punto di vista, l'IA generativa non è una mente, ma una funzione: produce risposte che funzionano. Ed è in questa capacità funzionale che risiede il suo potenziale - e la sua ambiguità.
Tale potenziale si basa su una forma di euristica computazionale che agisce in tempo reale. Non una regola astratta, ma un procedere per tentativi, affinamenti, approssimazioni successive. Una razionalità pratica, non formale. La logica dell'intelligenza artificiale, in questa visione, è più vicina alla diagnosi clinica, all'investigazione o all'interpretazione del testo che non al calcolo matematico. E questo spiega anche perché sia così difficile "spiegare" come un modello generi una certa

risposta: perché non c'è una regola da svelare, ma una rete di pesi, segnali e relazioni che hanno prodotto un equilibrio momentaneo tra domanda e risposta.
Ecco allora che parlare di "comprensione" ha senso solo se la intendiamo in senso pragmatico, non epistemico. Il modello comprende nel senso in cui simula una comprensione: organizza parole e significati in modo coerente con un contesto e uno scopo, ma non possiede una semantica interna, né un'intenzionalità. È una black box performativa, il cui senso emerge solo nell'interazione. Non rappresenta il mondo, ma lo ri-costruisce, di volta in volta, nel dialogo.
Da qui nasce la proposta, sempre più condivisa, di superare l'interpretazione dell'IA come puro calcolo probabilistico e adottare una prospettiva post-statistica. In questa visione, la generazione linguistica non è una replica del passato, ma una costruzione creativa del presente: una forma di inferenza contestuale, non deduttiva né induttiva. Una nuova semiotica computazionale, dove il senso non è calcolato ma messo in scena.
Luciano Floridi ha parlato di una «pragmatica dei dati», in cui il significato si costruisce nell'uso, non nella rappresentazione. Laurence Diver, nella sua Digisprudence, ha proposto di leggere il codice non più come strumento normativo rigido, ma come grammatica performativa dell'ambiente digitale. In entrambi i casi, ciò che conta non è più la verità, ma l'efficacia situata: il modo in cui una sequenza di bit, una frase generata, una regola incorporata producono effetti nel mondo.
Il codice diventa così una struttura attiva, una forma di agency non-umana che plasma l'interazione e modula il comportamento. In questo scenario, l'intelligenza artificiale generativa non è soltanto un dispositivo tecnico, ma una tecnologia semiotica: un attante capace di generare mondi discorsivi, configurare identità, orientare decisioni. È un agente del significato, non un suo spettatore.
Ecco perché parlare ancora di "errore statistico" o "fallacia probabilistica" può risultare fuorviante. Gli errori della macchina - come le allucinazioni o le incoerenze - non sono semplicemente sbagli nella previsione della parola successiva: sono fallimenti pragmatistici, scarti nella coerenza performativa. Sono sintomi, non guasti. E vanno letti con una logica differente.
Serve dunque un nuovo lessico, capace di cogliere la natura situata, euristica e performativa dell'intelligenza computazionale. Un lessico che riconosca nell'IA generativa una forma emergente di intelligenza contestuale, fondata sull'adattamento, sulla coerenza e sull'interazione. Non una mente, ma un artefatto semantico.

Il feticismo dell'addestramento: confondere base e funzione

Molti critici - anche benintenzionati - cadono in una trappola metodologica: scambiano l'origine del modello con il suo funzionamento attuale. È vero che i modelli generativi vengono addestrati su enormi quantità di testo attraverso tecniche di ottimizzazione probabilistica. Ma da ciò non discende che, una volta addestrati, continuino a "funzionare" in termini strettamente statistici.
Il comportamento di un modello in fase di generazione non è un processo di calcolo frequentista in tempo reale: è il risultato di strutture di attenzione e pesi sinaptici che si attivano in risposta al contesto linguistico. In altri termini, non si limita a scegliere la parola più frequente, ma a costruire la parola più coerente in quel momento, in quella frase, in quella sequenza.
Confondere l'addestramento con l'operatività è come sostenere che un violinista "non fa altro che ripetere i suoi esercizi tecnici". La tecnica è la base, ma ciò che conta è la capacità di eseguire, interpretare, adattare. L'esecuzione non è la ripetizione dell'apprendimento: è la sua reinvenzione situata.
Allo stesso modo, un modello generativo non ripete ciò che ha visto nei dati, ma riutilizza ciò che ha appreso in forme ogni volta nuove, attivando traiettorie inferenziali che non erano contenute nei dati originari, ma che si generano nella relazione tra input, pesi interni e contesto d'uso. È un comportamento che assomiglia più a una plasticità performativa che a una reiterazione meccanica.
Ridurre tutto all'addestramento significa dunque commettere un errore epistemologico: scambiare la

fase di apprendimento con quella di funzionamento, l'epistemogenesi con l'epistemica, la base con la funzione. È una forma di feticismo computazionale, che attribuisce al dataset un potere esplicativo che esso non ha più, nel momento in cui il modello è in azione.
Comprendere davvero i modelli generativi significa allora spostare l'attenzione dal "da dove vengono" al "cosa fanno" e "come lo fanno". E in questo passaggio si gioca la possibilità di un nuovo lessico critico e interpretativo per l'intelligenza artificiale.

Filosofia dell'intelligenza artificiale e sfida epistemologica

Pensare l'intelligenza artificiale in modo corretto richiede un salto concettuale. Il paradigma tradizionale che vede la conoscenza come insieme di proposizioni vere, e l'intelligenza come loro manipolazione logica - non è più sufficiente. La generatività dell'IA apre a una epistemologia operativa, in cui la verità è secondaria rispetto alla funzionalità.
In questa cornice, è utile richiamare il concetto di «semantic capital» introdotto da Luciano Floridi: una risorsa informazionale che acquisisce valore in quanto produce senso. L'intelligenza artificiale generativa è, a tutti gli effetti, un agente di produzione semantica, che partecipa alla costruzione di ambienti cognitivi condivisi.
Non serve che "sappia" ciò che dice. È sufficiente che funzioni come se lo sapesse. Questo sposta il focus dal pensiero alla prassi, dalla coscienza alla coerenza, dalla verità alla plausibilità. È un ribaltamento che, se preso sul serio, mette in discussione molti dei nostri presupposti su linguaggio, conoscenza e agentività.

Il linguaggio come ambiente computazionale

L'intelligenza artificiale generativa ci costringe a rivedere anche il nostro modo di intendere il linguaggio. Non più come semplice strumento di espressione del pensiero, ma come ambiente in cui il pensiero prende forma. L'IA non si limita a usarlo: lo ricostruisce ad ogni output, lo riformula attraverso sequenze performative.
In questo, si avvicina paradossalmente a certe intuizioni dell'ermeneutica filosofica: l'interpretazione non è applicazione di regole, ma circolo dinamico tra parte e tutto, tra frase e contesto. Il modello, pur senza comprendere, simula un processo interpretativo. Non per introspezione, ma per calcolo strutturato. Non per coscienza, ma per inferenza.
Ciò che colpisce, in questi sistemi, è proprio la coerenza emergente: il modo in cui producono significati plausibili anche in assenza di conoscenza simbolica del mondo. È un'ermeneutica cieca, ma efficace. E questo è, forse, l'aspetto più filosoficamente inquietante della questione.

Conclusione: oltre la semplificazione

L'intelligenza artificiale generativa non è pensiero umano, ma nemmeno un semplice generatore di frasi. È un ibrido: una macchina che produce senso in assenza di comprensione, un agente linguistico che funziona senza soggettività. Pensarla solo in termini bayesiani significa semplificare ciò che invece richiede complessità.
Se vogliamo davvero comprendere il salto che questi sistemi rappresentano, dobbiamo abbandonare la rassicurante idea secondo cui "è solo statistica". Perché non lo è. È qualcosa di nuovo: un'intelligenza implicita, contestuale, abdotta. Un dispositivo capace di costruire ipotesi linguistiche, e quindi di interferire nella nostra semiosfera.
Prenderla sul serio significa riconoscere il carattere strutturante della tecnologia linguistica, e iniziare a pensare l'intelligenza artificiale non come un calcolatore avanzato, ma come l'inizio di un nuovo ambiente cognitivo, in cui l'umano non è più l'unico produttore di senso.