L'illusione della conoscenza nei chatbot, che non sanno di non sapere: cosa sbagliamo nel dibattito sull'intelligenza artificiale generale *di Walter Quattrociocchi**

Dopo che le intelligenze artificiali hanno superato il test di Turing, le conversazioni sui large language model devono tornare a interrogarsi sulla differenza fra la plausibilità statistica e il significato della vera conoscenza (Fonte: https://www.corriere.it/ 17 febbraio 2026)

La morte di Socrate di Jacques-Louis David

Nel dibattito pubblico sull'intelligenza artificiale la domanda sembra sempre la stessa: **siamo già arrivati all'Agi, cioè** l'intelligenza artificiale generale? È una domanda comprensibile, ma sempre più fuori fuoco. Mentre il discorso mediatico continua a inseguire benchmark, test di Turing e dimostrazioni spettacolari, nella discussione scientifica internazionale il terreno si sta già spostando altrove. Non più verso ciò che le macchine sanno fare, ma verso **che tipo di conoscenza producono** quando parlano con la nostra voce senza condividere la nostra responsabilità epistemica.

Negli ultimi mesi il confronto si è intensificato sulle pagine delle principali riviste scientifiche e nei dibattiti tra ricercatori che lavorano direttamente sul tema dell'**affidabilità dei sistemi generativi**. In un recente intervento scritto insieme allo scienziato cognitivo Gary Marcus abbiamo sostenuto che gran parte della narrativa sull'Agi nasce da una confusione di fondo tra **approssimazione**

statistica e intelligenza generale. Il punto non è negare i progressi dei modelli linguistici. Sarebbe ingenuo. Il punto è riconoscere che prestazioni elevate in contesti controllati **non equivalgono automaticamente a competenza generale**. I benchmark misurano abilità circoscritte e riducono la complessità del mondo reale a compiti chiusi. Funzionano bene per tracciare miglioramenti incrementali, molto meno per valutare la robustezza, la trasferibilità e il comportamento sotto condizioni di incertezza.

Per capire davvero cosa sta succedendo, forse conviene fare un passo indietro e tornare al **test di Turing**. L'idea era semplice: se una macchina riesce a sostenere una conversazione indistinguibile da quella umana, possiamo attribuirle una **forma di intelligenza**. Era una proposta brillante per il suo tempo, ma oggi mostra tutti i suoi limiti. Il test misura la **somiglianza del comportamento, non la natura dei processi** che lo generano. Già con **Eliza**, negli anni Sessanta, bastavano poche regole sintattiche per evocare l'**illusione di profondità** senza alcuna comprensione sottostante. Oggi, per molti versi, le condizioni per il superamento comportamentale del test di Turing si sono verificate. I modelli linguistici conversano con una fluidità che rende **sempre più difficile distinguere l'umano dal sintetico**. Ma proprio per questo diventa evidente che il criterio non basta più. Se sembrare intelligenti equivalesse a esserlo, avremmo risolto il problema mezzo secolo fa. La vera domanda non è se le macchine parlino come noi, ma che cosa accade quando la produzione linguistica viene separata dalla **responsabilità epistemica**.

Una parte della narrativa recente sostiene che, poiché **non esiste una definizione universale di intelligenza**, non sia possibile escludere che i modelli linguistici siano intelligenti in senso pieno. È un ragionamento circolare. Sappiamo con crescente precisione come funzionano questi sistemi: modelli probabilistici che apprendono distribuzioni linguistiche e generano sequenze plausibili. Confondere questo processo con capacità di giudizio o comprensione non rappresenta un avanzamento teorico, ma uno **slittamento semantico** che ridefinisce continuamente i termini del dibattito per sostenere una narrativa sempre più fragile.

È qui che si apre una linea di confine spesso invisibile nel discorso pubblico. Non tra entusiasti e scettici dell'intelligenza artificiale, ma **tra chi misura la conoscenza e chi misura soltanto la performance**. In uno **studio pubblicato su *Pnas***, *The Simulation of Judgment in LLMs*, abbiamo mostrato che esseri umani e modelli linguistici possono produrre valutazioni simili sull'affidabilità delle fonti informative seguendo però processi opposti: gli umani modulano il giudizio in base all'incertezza e ai costi dell'errore, mentre i modelli tendono a riempire i vuoti con **risposte plausibili anche quando l'evidenza è incompleta**. L'allineamento superficiale degli output nasconde quindi una divergenza profonda nei meccanismi di valutazione.

Un quadro teorico più ampio emerge nel nostro **recente lavoro** *Epistemological Fault Lines Between Human and Artificial Intelligence*, dove mostriamo perché **l'allineamento linguistico non equivale a un allineamento epistemico**. Per un essere umano, giudicare significa più che produrre una risposta plausibile: significa distinguere tra ciò che si sa e ciò che non si sa, pesare l'evidenza,

riconoscere l'incertezza e, soprattutto, assumersi il costo dell'errore. Un giudizio è un atto che implica responsabilità, perché è legato a credenze, ragioni e conseguenze.
Un modello linguistico opera in modo radicalmente diverso. Non «controlla» il mondo, non forma credenze, non confronta ipotesi con dati, non aggiorna convinzioni alla luce di una smentita. Genera testo stimando, passo dopo passo, la **continuazione più probabile** in base a un contesto linguistico, ottimizzando la coerenza e la plausibilità statistica rispetto ai pattern appresi dai dati. È un meccanismo potente, ma di natura diversa: produce risposte, senza possedere le condizioni che rendono una risposta **conoscenza**.
Da qui nasce la frattura: **la fluenza può simulare il giudizio**. E quando la plausibilità linguistica sostituisce la valutazione epistemica, l'output può sembrare affidabile anche nei casi in cui un umano si fermerebbe, ridurrebbe la fiducia, o chiederebbe ulteriori evidenze. Non è un dettaglio filosofico: è il punto tecnico e istituzionale che decide quanta autorità siamo disposti a delegare a sistemi che «sanno parlare» ma **non sanno, in senso pieno, quando non sanno**.
Questo elemento diventa difficilmente aggirabile se lo si colloca dentro la scommessa economica sui Llm. Il ritorno si fonda sull'idea che ci sarà la delega, ma la delega senza affidabilità non si può fare.
Perché, quando si comprende davvero come sono costruiti questi sistemi, diventa evidente che un Llm può automatizzare la produzione di risposte, ma **non le condizioni che rendono una risposta affidabile.**
Quello che manca tanto, infatti, è la comprensione degli attrezzi che alimentano **narrazioni al limite dell'astrologia**.
Il rischio, infatti, non è tanto l'eccesso di scetticismo quanto la proliferazione di cornici interpretative costruite **senza un ancoraggio empirico solido**. Molti discorsi sull'AI moltiplicano definizioni e categorie teoriche senza chiarire la natura dei sistemi di cui parlano. Quando l'oggetto dell'analisi resta opaco, il dibattito tende a diventare autoreferenziale: più parole, meno comprensione.
Si ripete spesso che avremo sempre più bisogno delle materie umanistiche. Forse la questione è diversa: ciò che l'AI sta rendendo visibile non è una rinascita del pensiero critico, ma la fragilità di approcci che hanno **sostituito l'analisi con la retorica**.
Quello che probabilmente cambierà davvero è altro. In un mondo dove la plausibilità linguistica si disancora sempre più dalla logica, torneranno a contare approcci più duri, più verificabili, meno indulgenti verso la «fuffa» performativa. Perché alla fine, **passata la stagione delle chiacchiere, i problemi restano**. E qualcuno dovrà capirli e risolverli davvero.
E quando succede questo, la differenza tra chi parla e chi sa fare diventa improvvisamente molto costosa.

Mentre si discutono scenari sempre più astratti, la questione centrale resta sullo sfondo: come valutare **l'affidabilità epistemica** di sistemi che producono linguaggio credibile ma non condividono le condizioni umane della responsabilità, dell'incertezza e della revisione dell'errore. In questo senso, il superamento simbolico del test di Turing non segna l'arrivo dell'intelligenza artificiale generale, ma **la fine di un certo modo di pensarla**. Non basta più chiedersi se una macchina sembri umana. Bisogna capire quali condizioni rendono una risposta affidabile, quando **un sistema sa riconoscere i propri limiti** e quale differenza esiste tra generare frasi plausibili e produrre conoscenza condivisibile.
La vera posta in gioco non è stabilire se l'Agi sia già arrivata, ma comprendere **come sta cambiando la natura della conoscenza** in un ambiente dove la produzione di testo è sempre più automatizzata. La linea di frattura del nostro tempo non passa tra chi crede o non crede nell'AI. Passa tra chi scambia la fluenza linguistica per comprensione e chi riconosce che la conoscenza richiede qualcosa di più della somiglianza con il linguaggio umano. Quando questa distinzione diventerà esplicita, il dibattito sull'intelligenza artificiale **smetterà di essere una gara di prestazioni** e tornerà a essere una riflessione su come produciamo, condividiamo e difendiamo ciò che chiamiamo conoscenza.

***** ***Walter Quattrociocchi è professore ordinario di Informatica all'Università La Sapienza di Roma***