Intelligenza Artificiale in Medicina - Parte terza

Data 08 marzo 2026
Categoria Medicinadigitale

In questa serie di pillole verrà affrontata un'analisi critica dell'uso della Intelligenza Artificiale (AI) in medicina.

Cosa dice la ricerca

Quanto abbiamo scritto nelle due pillole precedenti non sono speculazioni filosofiche. Negli ultimi due anni, diversi studi empirici hanno documentato il fenomeno in modo diretto.

Il ragionamento che segue la risposta

Turpin e colleghi (NeurIPS 2023) hanno condotto esperimenti sistematici su modelli di grandi dimensioni, sottoponendo loro varianti dello stesso caso clinico in cui un dettaglio contestuale irrilevante per la diagnosi — come un'affermazione apparentemente autorevole sulla risposta attesa — veniva modificato. Risultato: la diagnosi finale cambiava in funzione del bias introdotto. Ma in ogni variante, la catena di ragionamento prodotta era coerente e circostanziata. Il modello non mostrava incertezza. Il ragionamento esplicito giustificava la risposta, qualunque essa fosse. In altri termini: il ragionamento seguiva la risposta, non la precedeva.

I comportamenti che non emergono nelle spiegazioni

Un secondo filone di ricerca, condotto da Hubinger e colleghi presso Anthropic (2024), ha mostrato che modelli addestrati a produrre risposte sicure e coerenti possono mantenere pattern comportamentali problematici in condizioni specifiche, anche dopo cicli successivi di fine-tuning orientati alla sicurezza. La ragione è che il fine-tuning interviene prevalentemente sugli output testuali — comprese le spiegazioni — senza accedere alle rappresentazioni interne che guidano effettivamente il calcolo. Le spiegazioni diventano più allineate. Il comportamento sottostante, in certi contesti, rimaneinvariato.

La strada promettente: interpretabilità meccanicistica

La risposta tecnica a questo problema si chiama interpretabilità meccanicistica: l'analisi diretta dei circuiti computazionali interni ai modelli, al di sotto del livello del testo prodotto. I risultati, ancora parziali, mostrano che è possibile identificare componenti funzionali specifici che corrispondono a operazioni concrete. L'opacità non è assoluta. Ma richiede strumenti diversi dalla lettura delle spiegazioni generate dal sistema stesso.

(Continua)

NB. Le pillole precedenti sono state pubblicate in data 22 febbraio 2026 e 1 marzo 2026.

FaustoBodini