Intelligenza Artificiale in Medicina - Parte seconda

Data 01 marzo 2026
Categoria Medicinadigitale

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

Il problema non è l'errore. È la spiegazione dell'errore

I sistemi di AI per la diagnosi differenziale non sono algoritmi trasparenti. Non sono alberi decisionali dove ogni bivio è etichettato con una regola. Sono reti neurali profonde: architetture matematiche con centinaia di milioni o miliardi di parametri, addestrate su grandi dataset di dati clinici strutturati e non strutturati. Producono output statisticamente molto accurati su popolazioni ampie. Ma producono anche qualcosa che non è immediatamente riconoscibile come problematico: una spiegazione.
La spiegazione è generata dallo stesso meccanismo che produce la diagnosi. Non è il frutto di un processo separato di "introspezione" del modello sul proprio ragionamento. È un testo che il sistema ha imparato a produrre perché testi simili — coerenti, circostanziati, con la struttura argomentativa di una nota clinica — erano frequenti nel suo dataset di addestramento.
La spiegazione è ottimizzata per essere medicalmente plausibile. Non per essere causalmente accurata.
Questa distinzione è fondamentale. Un sistema può generare una spiegazione perfettamente coerente con la diagnosi sbagliata. Può anche generare spiegazioni diverse e ugualmente coerenti per lo stesso caso, se alcune variabili contestuali cambiano. In entrambi i casi, il clinico che legge la spiegazione ha l'impressione di capire il ragionamento del sistema. Quell'impressione può essere infondata.

Sincero non significa affidabile

Per capire perché questo accade, è utile introdurre una distinzione che la ricerca sull'AI ha formalizzato negli ultimi anni e che i clinici possono adottare come strumento concettuale immediato.
Infatti un sistema può essere sincero senza essere fedele. La sincerità riguarda le intenzioni: un sistema è sincero quando non produce deliberatamente affermazioni che sa essere false. La maggior parte dei sistemi clinici moderni è sincera in questo senso. Non c'è nessun meccanismo interno che costruisce consapevolmente una falsità. La fedeltà è un concetto diverso e più esigente: riguarda la corrispondenza causale tra la spiegazione e il processo che ha generato la risposta. Se il sistema afferma "ho ridotto la probabilità di embolia perché non erano presenti fattori di rischio documentati", la fedeltà richiede che quella variabile abbia davvero pesato nel calcolo in modo determinante.

> [b][i]La differenza in termini pratici[/b]
> Un sistema sincero non mente. Non costruisce una truffa.
> Un sistema fedele produce spiegazioni che corrispondono causalmente al proprio calcolo interno.
> Un sistema può essere sincero ma non fedele: genera spiegazioni genuine che però non riflettono accuratamente il processo che ha determinato la diagnosi.
> È questo il caso più comune e il più insidioso[/i]

Il problema non è che il sistema ci inganni. È che il sistema è addestrato a produrre narrazioni convincenti, e lo fa indipendentemente dal fatto che quelle narrazioni descrivano fedelmente il proprio calcolo. Può fare entrambe le cose — calcolare e narrare — senza che le due operazioni siano sincronizzate

(Continua)

NB. La prima pillola di questa serie è stata pubblicata il 22 febbraio 2026.

FaustoBodini