Intelligenza Artificiale in Medicina - Parte sesta

Data 29 marzo 2026
Categoria Medicinadigitale

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

Cosa può fare il clinico, oggi

Non è ragionevole — né utile — concludere che i sistemi di AI siano da evitare. La loro accuratezza media su casi tipici è documentata e spesso notevole. Il problema è l'uso acritico, e la soluzione non richiede competenze informatiche: richiede un'igiene cognitiva specifica.

1. Trattare la spiegazione come un'ipotesi diagnostica, non come una prova
La spiegazione del sistema è un punto di partenza per il ragionamento clinico, non la sua conclusione. Vale la pena chiedersi esplicitamente: questa spiegazione è compatibile con tutti i dati che ho, inclusi quelli che il sistema potrebbe non aver elaborato? Ci sono elementi nel quadro clinico che la spiegazione non menziona?

2. Cercare attivamente le alternative scartate
Quando un sistema assegna una probabilità dominante a una diagnosi, l'attenzione tende a convergere su quella. È più utile guardare la distribuzione completa e chiedersi: la seconda e la terza ipotesi sono state escluse sulla base di dati reali, o solo perché il sistema non aveva accesso alle informazioni rilevanti?

3. Non sopravvalutare la coerenza narrativa
Una spiegazione ben costruita non è garanzia di un calcolo accurato. La fluidità del testo e la sua aderenza al registro medico sono proprietà del linguaggio, non dell'inferenza diagnostica. Un sistema può scrivere una spiegazione impeccabile per una diagnosi sbagliata.

4. Applicare più scetticismo dove l'AI sembra più sicura
I casi in cui la spiegazione è più fluida e la probabilità più alta sono spesso quelli più tipici e frequenti nel dataset di addestramento. I pazienti atipici — presentazioni rare, comorbidità complesse, dati parziali — sono quelli su cui la dissociazione tra spiegazione e processo è più probabile.

5. Documentare il proprio ragionamento separatamente
Se la spiegazione del sistema entra nella documentazione clinica, è buona pratica — e, con l'AI Act, una responsabilità crescente del deployer — annotare esplicitamente se e perché se ne è tenuto conto, quali elementi aggiuntivi sono stati considerati, e dove il giudizio clinico ha divergito dal suggerimento del sistema. Non è burocrazia: è la traccia di ciò che il sistema non può fare.

6. Partecipare ai processi di valutazione del vendor
L'AI Act obbliga i deployer a cooperare con i fornitori nel monitoraggio post-deployment. Questo significa che le segnalazioni di comportamenti anomali del sistema — spiegazioni incoerenti, output sorprendenti, diagnosi discordanti con il quadro clinico — non sono solo episodi da ignorare, ma dati preziosi per la sorveglianza continua del sistema. I clinici sono, di fatto, una parte del meccanismo di supervisione previsto dal Regolamento.

La prossima frontiera: certificare le spiegazioni, non solo le diagnosi

Il problema della dissociazione tra spiegazione e calcolo non si risolve rendendo i sistemi più verbosi o più articolati nelle loro giustificazioni. Si affronta a un livello diverso: quello delle rappresentazioni interne.
La ricerca in interpretabilità meccanicistica sta sviluppando tecniche — il probing sui residual stream, il causal tracing, l'activation patching — che permettono di verificare se una specifica variabile ha davvero influenzato causalmente la risposta del modello, indipendentemente da ciò che il sistema afferma nella sua spiegazione. Non sono ancora strumenti clinicamente operativi. Ma indicano la direzione.
L'obiettivo pratico, coerente con lo spirito dell'AI Act e con la direzione della ricerca, è quello di sistemi certificati non solo per l'accuratezza diagnostica su popolazioni di test, ma per la fedeltà verificata delle proprie spiegazioni: sistemi in cui la trasparenza non sia una narrativa prodotta per rassicurare il clinico, ma una proprietà misurabile, auditabile e monitorata nel tempo.
La distinzione tra ciò che il sistema "dice" di aver calcolato e ciò che ha effettivamente calcolato rimane, per ora, una responsabilità del clinico. Non è un onere irragionevole: è la stessa responsabilità che si esercita ogni volta che si interpreta un referto redatto da chi non ha visto il paziente, si valuta una consulenza telefonica, si pesa un risultato di laboratorio nel contesto di un quadro clinico che il laboratorio non conosce.

[i] [b]Cinque punti da ricordare[/b]

1. I sistemi di AI per la diagnosi differenziale producono spiegazioni che possono non corrispondere al processo computazionale reale.

2. Un sistema può essere sincero — non ingannevole — senza essere fedele: la spiegazione è plausibile, ma non necessariamente causale.

3. Il rischio è maggiore nei casi atipici e con dati parziali: esattamente quelli in cui il supporto AI sembra più utile.

4. L'AI Act (Reg. UE 2024/1689) classifica questi sistemi come ad alto rischio e impone obblighi di trasparenza, supervisione umana e alfabetizzazione AI al deployer — cioè alla struttura e al clinico.

5. Documentare il proprio ragionamento separatamente da quello del sistema non è burocrazia: è la traccia di ciò che l'AI non può fare e un obbligo emergente ai sensi del Regolamento europeo.[/i]

Riferimenti e note

Turpin M. et al. "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting." NeurIPS 2023.

Hubinger E. et al. "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training." Anthropic, 2024.

Jain S., Wallace B.C. "Attention is not Explanation." NAACL-HLT 2019.

Meng K. et al. "Locating and Editing Factual Associations in GPT." NeurIPS 2022.

Olah C. et al. Mechanistic Interpretability research, Anthropic / Distill, 2020–2024.

Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio, del 13 giugno 2024 (AI Act), GU L 2024/1689, 12.7.2024. Articoli 13 (trasparenza), 14 (supervisione umana), 26 (obblighi dei deployer).

FDA: "Artificial Intelligence/Machine Learning (AI/ML)

Based Software as a Medical Device (SaMD) Action Plan," 2021.

Il caso clinico di apertura è uno scenario costruito a scopo illustrativo, fedele a dinamiche documentate in letteratura ma non riferito a eventi reali.

(fine)

NB. Le pillole precedenti di questa serie sono state pubblicate in data 22 febbraio 2026, 1 marzo 2026, 8 marzo 2026, 15 marzo 2026, 22 marzo 2026.

FaustoBodini