Quanto sono attendibili i chatbot AI che danno consigli medici? di Ruggiero Corcella

Da un lato, esistono segnali incoraggianti: in contesti sperimentali, alcuni modelli hanno dimostrato di poter raggiungere livelli di accuratezza paragonabili a quelli dei medici. Dall'altro, manca però una validazione indipendente (Fonte: https://www.corriere.it/ 2 maggio 2026)

Quanto sono attendibili i chatbot AI? **Secondo un'analisi di MIT Technology Review, la risposta è ancora incerta**. Da un lato, esistono segnali incoraggianti: alcuni modelli hanno dimostrato di poter fornire consigli utili e, in contesti sperimentali, di raggiungere livelli di accuratezza paragonabili a quelli dei medici. Dall'altro, **gli esperti sottolineano una mancanza cruciale: la validazione indipendente**. Questo rappresenta un problema soprattutto in ambito sanitario, dove errori o imprecisioni possono avere conseguenze gravi. **Uno dei punti più critici è il cosiddetto triage,** cioè la capacità di distinguere tra situazioni urgenti e non urgenti.

Crash test fallito

«Un recente lavoro pubblicato su *Nature Medicine* dimostra come il "crash test" non sia andato bene — sottolinea Simona Giubilato, coordinatrice dell'Area dedicata all'Intelligenza Artificiale in Cardiologia di Anmco —. In questo studio ricercatori statunitensi **hanno esaminato l'affidabilità di ChatGPT Salute**, lanciato a gennaio da OpenAI ma non ancora disponibile in Italia. Ebbene, su quattro possibili suggerimenti da dare che andavano dal restare a casa al recarsi subito al **Pronto Soccorso** l'applicazione è risultata molto accurata nei casi intermedi ma **ha sopravvalutato sintomi trascurabili nel 65 per cento dei casi. E, soprattutto, ha sottovalutato emergenze gravi**

in più del 50% dei casi. In breve, consigliava a pazienti che avrebbero dovuto correre in ospedale di aspettare un giorno o due per farsi visitare dal medico di famiglia».

Risposte «problematiche»

Uno studio pubblicato su BMJ Open ha valutato **l'affidabilità di cinque chatbot di intelligenza artificiale (Gemini, DeepSeek, Meta AI, ChatGPT e Grok) su temi di salute.** I risultati mostrano che il 50% delle risposte è problematico: il 30% parzialmente e il 20% gravemente, con rischi di disinformazione e possibili danni per gli utenti. Le prestazioni sono simili tra i sistemi, ma **Grok ha prodotto più risposte critiche**, mentre Gemini meno. **I chatbot risultano più accurati su vaccini e cancro, meno su nutrizione, cellule staminali e performance atletiche**. Le risposte, spesso sicure ma poco supportate da fonti affidabili, presentano anche citazioni errate e alta complessità linguistica. Lo studio evidenzia limiti strutturali dell'IA, che non valuta le prove ma genera testi probabilistici. I ricercatori chiedono maggiore educazione pubblica, formazione e regolamentazione per evitare che l'uso crescente dei chatbot comprometta la salute pubblica.

Le «allucinazioni»

Per non parlare delle «allucinazioni»: «Il dato più allarmante viene da **un trial randomizzato su studenti di medicina pubblicato nel 2026.** In questo studio le spiegazioni fuorvianti fornite dall'AI hanno ridotto significativamente l'accuratezza diagnostica, mentre le spiegazioni corrette non hanno offerto miglioramenti significativi. **La disinformazione si è dimostrata più potente e robusta nel causare danni,** di quanto le informazioni corrette lo siano nel produrre benefici. E in certi contesti un falso negativo non segnalato è più pericoloso di un errore riconoscibile», aggiunge.

A rischio di hackeraggio

Non basta. Uno studio pubblicato su JAMA Network Open ha dimostrato che **i chatbot «medici» possono anche essere manipolati con estrema facilità**. In simulazioni controllate, attacchi informatici mirati sono riusciti a modificare le risposte dei chatbot nel 94,4% dei casi, inducendo raccomandazioni cliniche errate o pericolose.

Le «black box»

L'AI sa arrivare alla diagnosi giusta, ma spesso non sa spiegare come. Uno studio del Mass General Brigham (Usa), pubblicato su JAMA NetworkOpen, **ha testato 21 chatbot , inclusi i modelli più recenti di ChatGPT, DeepSeek, Claude, Gemini e Grok** superano il 90% di accuratezza con dati completi, ma falliscono oltre l'80% nelle diagnosi differenziali.

La trappola delle malattie false: la bixonimania

Occhi irritati e pruriginosi? Sintomi molto comuni, se si passa troppo tempo a fissare gli schermi. Se poi ci si strofina troppo gli occhi, le palpebre potrebbero assumere una leggera sfumatura rosata.

Fin qui, tutto normale. Ma se, negli ultimi 18 mesi, aveste digitato questi sintomi in una serie di chatbot popolari chiedendo cosa non andasse in voi, avreste potuto ricevere una risposta insolita: **bixonimania. Una falsa malattia, inventata da un team guidato da Almira Osmanovic Thunström**, una ricercatrice medica dell'Università di Göteborg, in Svezia, pubblicando studi fittizi per testare i limiti dei modelli linguistici. L'esperimento, come riferisce **Nature**, ha funzionato: in poche settimane **i principali chatbot hanno descritto la patologia come reale**, arrivando a suggerire diagnosi e cure. Ancora più preoccupante, i falsi studi sono stati citati anche in articoli scientifici peer-reviewed, segno di un uso superficiale delle fonti.