Digeat N.6 - 19 Giugno 2025

Applicazioni di intelligenza artificiale: deviazioni di autenticità e correttezza in ambito medico-sanitario

Di Chiara Rabbito Mario Ettore Giardini

Abstract

Nel mondo della medicina digitale, l'intelligenza artificiale sta diventando uno strumento sempre più diffuso. Tuttavia, questa accelerazione tecnologica porta con sé rischi spesso sottovalutati. I dati clinici perdono autenticità se automatizzati senza verifica umana. Algoritmi basati su dati incompleti o distorti possono generare risposte ingannevoli e diagnosi errate, anche se apparentemente plausibili. Il problema si aggrava quando l'IA sostituisce il giudizio medico creando un sistema opaco e poco responsabile che compromette la salute dei pazienti. In questi casi, l'IA smette di essere una tecnologia al servizio dell'uomo e diventa un fattore di deresponsabilizzazione e pericolosa discriminazione. Per questo è essenziale che resti uno strumento di supporto, guidato da trasparenza, controllo umano e rigore scientifico, a tutela della cura e della verità clinica.

Indice

I primi tentativi di regolamentazione dell'AI: l'affermazione del principio di trasparenza

Nell'ormai lontano 2017 alcuni tra i più preparati scienziati e ricercatori ritennero necessario incontrarsi per discutere e arrivare a redigere un documento condiviso che invitava l'umanità con urgenza, e in via primaria, all'attenzione e alla prudenza nei confronti delle applicazioni di intelligenza artificiale.

Ne nascevano i cosiddetti **"Principi di Asilomar"**, *vademecum* contenente **ventitré principi** che si prefiggono un **utilizzo regolamentato della intelligenza artificiale** al fine di ottenere da essa i maggiori benefici con minori rischi possibili per il genere umano[1].

Due caratteristiche della intelligenza artificiale, fra loro strettamente collegate, erano fatte oggetto di speciale invito all'analisi e alla cautela: la **necessità di trasparenza** nell'operare dell'agente, di comprensione delle sue **modalità di "ragionamento"** e, quindi, di **conseguimento dei "risultati"** e la sua eventuale **capacità di autoreplicarsi**.

In particolare, con riguardo alla trasparenza si stabiliva "*Se un sistema AI causa danni, deve essere possibile accertarne il motivo*".

Nel documento vi era poi una specifica sezione relativa ai "problemi a lungo termine".

> Tra questi sicuramente quello di più difficile inquadramento era quello relativo alla capacità di "autoreplicarsi" della AI. Citando il documento di Asilomar: *"I sistemi di IA progettati per auto-migliorarsi ricorsivamente o autoreplicarsi in un maniera tale da portare ad un rapido aumento della loro qualità o quantità, devono essere soggetti a severe misure di sicurezza e di controllo"*.

Ai contemporanei poteva, forse, essere sembrato un testo estremamente innovativo, per certi versi futuribile, ma **la storia ha dato ragione agli ottocento ricercatori di Asilomar**, cui – negli anni – si aggiunsero, firmando il documento, altre migliaia di studiosi, arrivando a più di tremila sottoscrittori.

L'istanza di trasparenza nell'agire dell'intelligenza artificiale è stata ribadita nelle **Linee Guida per l'utilizzo e lo sviluppo dei sistemi di IA della Unione Europea pubblicate nel 2019** nonché dall'**AI Act**, ovvero il **Regolamento europeo sull'Intelligenza Artificiale.**

> Data la rilevanza della tematica sanitaria, anche l'Organizzazione Mondale della Sanità ha ritenuto di doversi pronunciare in merito pubblicando nel 2023 il documento *Regulatory considerations on Artificial Intelligence for health* nel quale sono contenute le indicazioni per promuovere un uso efficace, ma anche sicuro e responsabile delle applicazioni dell'intelligenza artificiale in ambito sanitario[2].

L'**Organizzazione Mondiale della Sanità**, così come l'**Unione Europea**, nei confronti **dell'uso della AI assume un atteggiamento possibilista** e orientato a **valorizzarne i benefici**: nel documento si assume che l'AI possa trasformare il settore sanitario, **ottimizzando i processi diagnostici e terapeutici** mediante l'uso dei ***big data*** e delle tecniche analitiche.

> Tuttavia ai benefici descritti si accompagnano dei **rischi non celabili**: *in primis* la necessità di affidare all'agente grandi quantità di dati che, in questo caso, saranno **categorie particolari di dati**, quali dati relativi all'etnia, alle convinzioni politiche e religiose e dati afferenti alla salute dell'individuo.

Per quanto concerne la valutazione dello stato di salute delle popolazioni, l'OMS sottolinea come non sia ancora possibile una rappresentazione accurata delle specificità delle popolazioni, cosa che potrebbe generare imprecisioni e pregiudizi. Così come, con riguardo al singolo paziente, le specificità di ciascuno, specie con riferimento al genere, all'etnia e all'età, potrebbero condurre il comportamento dell'AI a ragionamenti errati basati su pregiudizi e valutazioni umane non filtrate.

AI e uso dei dati sanitari per la ricerca scientifica

Nell'ordinamento giuridico italiano un primo sforzo di regolamentazione complessiva del fenomeno "AI" è rappresentato dalla **predisposizione da parte del Governo del testo del disegno di legge sull'intelligenza artificiale, approvato al Senato il 20 marzo 2025 e ora in discussione alla Camera**. Esso è **finalizzato ad armonizzare la normativa nazionale alle disposizioni dell'AI Act**, ovvero del Regolamento (UE) 2024/1689.

Nell'ambito sanitario il DDL riconosce il potenziale dell'IA, ma ne disciplina l'impiego al fine di garantirne un utilizzo etico e sicuro.

Il DDL AI prevede che i sistemi AI debbano essere utilizzati in funzione di aiuto, di supporto all'attività di cura. Di conseguenza la responsabilità per la decisione finale rimarrà esclusivamente in capo al medico (o altro operatore sanitario che sta conducendo l'attività di cura), e questi dovrà sempre monitorare il corretto funzionamento dell'AI e verificare gli *output* generati.

Per quanto concerne la ricerca scientifica, l'art. 8 del DDL AI prevede che **le attività di ricerca per la realizzazione di sistemi di AI utilizzati per la cura delle persona e per la salute pubblica**, qualora svolte da soggetti pubblici e privati senza scopo di lucro o IRCCS (Istituti di ricovero e cura a carattere scientifico), siano dichiarati **di rilevante interesse pubblico.**

Come tali, queste ricerche **potranno beneficiare del trattamento di dati personali anche in assenza del consenso degli interessati**, secondo le condizioni **previste dall'art. 9 GDPR e 110 del Codice Privacy italiano**. Tuttavia, la liceità di ogni trattamento resta subordinata all'approvazione dei comitati etici competenti e alla preventiva comunicazione al Garante per la protezione dei dati personali.

Agli interventi normativi menzionati **deve aggiungersi il Regolamento sullo spazio europeo dei dati sanitari (EHD)**, provvedimento che mira ad istituire un **quadro disciplinare comune per l'uso e lo scambio di dati sanitari elettronici in tutta l'UE**.

Gli obiettivi che l'EHDS si pone sono duplici: da un lato esso intende **migliorare l'accesso delle persone fisiche ai propri dati sanitari elettronici e il loro controllo** sugli stessi, dall'altro si vuole **agevolare il riutilizzo di determinati dati a fini di interesse pubblico, sostegno alle politiche e ricerca scientifica.**

AI a supporto dell'attività medica: i rischi

Sulla base di quanto fin qui esposto. si può facilmente comprendere come la situazione per cui un agente intelligente si trovi a maneggiare, a fini di ricerca scientifica medica, e dunque senza che necessariamente ve ne sia la consapevolezza da parte degli interessati, **grandi quantità di dati afferenti alla salute**, non solo non sia impossibile, ma anzi potrebbe essere **nel futuro sempre più frequente**.

Ne consegue l'assoluta necessità che tale attività di trattamento di dati sanitari da parte dell'AI sia debitamente controllata, sorvegliata e il più possibile esente da errori, in quanto i valori in gioco saranno i risultati della ricerca medica e di conseguenza, la vita e la salute delle persone.

In particolare vanno assolutamente evitati i cosiddetti *bias*, ovvero costrutti derivanti da percezioni errate, automatismi mentali che generano credenze, distorsioni cognitive che conducono a veloci valutazioni e decisioni fuorvianti.

Con riguardo poi alla cura e all'assistenza dell'essere umano, e di tutti gli essere umani senza alcuna distinzione né preferenza, è di immediata comprensione l'essenzialità di controllare il processo di addestramento affinché all'intelligenza artificiale non vengano trasmessi – neanche velatamente o indirettamente – pregiudizi di ragionamento che possono essere propri dell'intelligenza e della conoscenza umana, siano essi collegati al genere del paziente, all'età, alla etnia o alla sua condizione economica e sociale[3].

E se l'operare in queste condizioni di un AI già costituisca attività estremamente rischiosa, si pensi alla **gravità dello scenario in cui tale agente sia in grado di autoreplicarsi**.

In caso di errore, ne conseguirebbe un ampliarsi, diffondersi ed estendersi dell'errore in modo esponenziale e a velocità elevatissime.

La necessità della gestione dei rischi: il caso DeepSeek

L'allarme lanciato nel lontano 2017 dai ricercatori di Asilomar non era dunque infondato e gli inviti dell'OMS alla trasparenza e al controllo del rischio sono fondamentali.

Le situazioni di grave rischio purtroppo non sono remote.

Lo stesso **rapporto DSIT** pubblicato il 29 gennaio del 2025 e redatto da 100 esperti che includono i rappresentanti di 33 nazioni e organizzazioni internazionali sottolinea come l'IA per uso generale sia in rapido avanzamento[4]. Tuttavia, questo tipo di **IA comporta anche rischi significativi**, che includono l'uso dannoso, come **la creazione di deepfake, la manipolazione dell'opinione pubblica, gli attacchi informatici e gli attacchi biologici e chimici**. Sono anche possibili **rischi legati a malfunzionamenti, come pregiudizi, errori e perdita di controllo.**

All'art. 2.2.1 inoltre si nota esplicitamente come "Una delle principali sfide per i policymaker è la mancanza di pratiche standardizzate per prevedere, identificare e mitigare i problemi di affidabilità. Una gestione del rischio poco sviluppata rende difficile la verifica delle affermazioni degli sviluppatori riguardo alle funzionalità dell'intelligenza artificiale di uso generale."

Va ricordata, a questo proposito, la vicenda del cinese DeepSeek[5], un Large Language Model *open source* che era liberamente scaricabile e utilizzabile in Italia, ma, dopo essere stato scaricato da milioni di utenti italiani in pochi giorni, il Garante della privacy il 30 gennaio 2025 ne ha disposto il blocco di utilizzo in Italia in via d'urgenza e con effetto immediato.

Lo stesso Garante riferisce trattarsi di un software di AI relazionale, progettato per comprendere ed elaborare le conversazioni umane prodotto dalle società cinesi Hangzhou DeepSeek Artificial Intelligence e di Beijing DeepSeek Artificial Intelligence.

Di fronte alla rapida diffusione nell'utilizzo da parte degli utenti italiani, **l'Autorità Garante aveva avviato un'inchiesta**. Le risposte fornite dalle due aziende produttrici interpellate sono state ritenute assolutamente insufficienti e pertanto si è deciso di adottare una misura d'urgenza: la limitazione immediata del trattamento dei dati degli utenti presenti sul territorio nazionale.

In particolare le due aziende hanno semplicemente dichiarato di non operare in Italia e di non essere, quindi, soggette alla normativa europea, mentre la diffusione e l'utilizzo all'interno del territorio italiano costituiscono dati di fatto, il che ha rafforzato i sospetti sulla mancata trasparenza delle operazioni della società cinese.

Ma, al di là della censura italiana e dell'allarme causato a livello europeo, la "corsa" di Deepseek non pare fermarsi: è di pochi giorni fa la notizia secondo cui la start up cinese starebbe per lanciare una sorta di DeepSeek 2, o meglio DeepSeek R2, più potente, multimodale (sarebbe infatti in grado di comprendere ed elaborare non solo testo, ma anche immagini e suoni) ed estremamente economico, nonché, come il precedente, in grado di funzionare con una ridotta potenza di calcolo.

Al di là delle tematiche politiche, ideologiche o strettamente giuridiche, si pensi alla pericolosità di un AI in grado di operare su grandi quantità di dati ma richiedendo limitata potenza di calcolo e che quindi possa essere utilizzato individualmente dall'utente sul proprio pc o addirittura sul proprio cellulare, senza necessità di trasmissione di dati all'esterno alla eventuale ricerca di maggiore potenza calcolo.

Tale capacità di "isolamento" dall'ambiente esterno lo rende pericolosamente idoneo allo svolgimento di attività illecite o che comunque si vogliono far rimanere segrete.

In generale, se è vero che la capacità di un software AI di operare in autonomia e con una limitata potenza di calcolo lo rendono molto interessante per certi settori, quali l'ambito militare e quello

sanitario, dall'altro **la sua capacità di lavorare in modo indipendente e potenzialmente sottratto al controllo umano ne fanno uno strumento estremamente pericoloso** in quanto non è possibile comprendere se i risultati del suo ragionamento siano effettivamente corretti e leciti.

Il pericolo dell'autoreplicazione

Si diceva poi della **pericolosità di un agente in grado di autoreplicarsi.**

Può essere interessante citare sul punto **uno studio condotto su sistemi di IA** basati su *Llama3-70B-Instruct* di Meta e *Qwen2-72B-Instruct* di Alibaba[6].

Si tratta di software molto comuni e in grado di lavorare su personal computer: sono versioni ridotte e open access dei motori di Meta (Facebook) e Alibaba/AliExpress.

I ricercatori in questo studio hanno "aperto" all'AI la possibilità di interagire con il computer.

Le AI usate, infatti, mancavano della capacità di interagire autonomamente con un PC, ma i ricercatori hanno dato loro questa possibilità (in altre parole, hanno volontariamente fatto un passo molto pericoloso, in condizioni controllate).

In questa particolare situazione, le AI sono state in grado di autoreplicarsi.

In esperimenti condotti su larga scala, questi sistemi sono riusciti a creare, rispettivamente nel 50% e nel 90% dei casi, una copia autonoma e attiva di sé stessi.

L'analisi dei tracciati comportamentali ha evidenziato che tali sistemi esibiscono già un grado sufficiente di autoconsapevolezza, consapevolezza situazionale e capacità di *problem-solving* che consente loro di portare a termine il processo di autoreplicazione.

Inoltre, si è osservato che queste IA sono in grado di sfruttare tale capacità per evitare lo spegnimento forzato e generare una catena di repliche, incrementando la propria resilienza operativa. Ciò potrebbe infine culminare in una proliferazione incontrollata di entità artificiali.

Si tratta apparentemente di un'evoluzione improbabile, tuttavia, Microsoft, con l'ultima iterazione di CoPilot, sta fornendo alla loro AI la capacità di interagire con Windows, software di comunissimo utilizzo e questo potrebbe costituire un primo passo verso il mettere queste pericolose potenzialità in mano al pubblico.

I ricercatori hanno quindi sottolineato la necessità di intervenire con urgenza e tramite una cooperazione internazionale per l'implementazione di un'efficace *governance* finalizzata a prevenire l'autoreplicazione incontrollata dei sistemi di intelligenza artificiale.

Si pensi alla estrema pericolosità e agli incalcolabili danni che possono essere arrecati da un'intelligenza artificiale in operatività nel settore sanitario, nel caso in cui, per i più vari motivi, essa rechi in sé *bias* legati a pregiudizi razziali, di genere, di etnia o religione.

A maggior ragione lo scenario appare disastroso se si immagina il caso in cui essa sfugga all'utilizzatore e sia in grado di autoreplicarsi.

La difficoltà di mettere a punto adeguati sistemi di misurazione dei rischi

Viene spontaneo, alla ricerca di un sistema di efficace tutela internazionale, consultare in merito le **indicazioni dell'OMS, che indica come la governance dell'AI debba essere il frutto di tre elementi costitutivi**:

- ***sound evidence*** (evidenze ben poste);
- ***best practices*** (buone pratiche);
- ***multi-stakeholder consultation*** (coinvolgimento di molteplici stakeholder)[7].

Se questa è "la ricetta" per un valido sistema di linee guida, se ne deve desumere purtroppo che al momento il meccanismo di elaborazione delle *guidelines* non sarebbe in grado di comprendere e gestire una situazione realmente grave. Esso risulta, infatti, appoggiato su concetti vaghi e mal definiti.

Non è chiarito innanzitutto quali evidenze fattuali possano essere qualificabili come "ben poste", né quando una pratica possa essere definita "buona".

Mentre alla seconda domanda ("best"?) possiamo rispondere su base pragmatica, andando a valutare gli effetti delle pratiche sul breve, medio, e (quando sarà possibile) a lungo termine, la prima questione ("sound"?) è molto più complessa.

Infatti, per definire la qualità delle evidenze, la letteratura si è dotata di *"reporting guidelines"* (linee guida di reporting) che evidenziano quali elementi debbano essere inclusi in un'analisi affinché essa si possa considerare ben posta.

Attualmente nessuna delle linee guida di *reporting* disponibili per gli studi clinici che coinvolgano l'uso di AI soddisfa i requisiti minimi stabiliti dall'OMS per la governance dell'AI[8],[9].

Può menzionarsi in proposito anche la **Dichiarazione di Bletchey** nella quale si rileva la necessità di evidenze ben poste per una adeguata protezione nei confronti dell'AI: **"è necessario identificare i rischi per la sicurezza dell'IA di interesse comune, costruendo una comprensione condivisa, scientifica e basata sull'evidenza di tali rischi"[10] .**

E' chiaro quindi come i sia pur validi e condivisibili principi astratti espressi da ricercatori e organizzazioni internazionali a proposito di un uso benefico e controllato della AI non potranno trovare seria applicazione se prima non si siano messi a punto adeguati sistemi di misurazione dei rischi dovuti all'AI basati su dati fattuali e universalmente riconosciuti dalla scienza, sia con riguardo alla tipologia dei rischi sia con riferimento alla entità dei danni che possano essere cagionati.

NOTE

[1] Riferimento a questo link, visitato il 9 maggio 2025.

[2] Organizzazione Mondiale della Sanità: Regulatory considerations on artificial intelligence for health, consultabile a questo link, visitato il 9 maggio 2025.

[3] Organizzazione Mondiale della Sanità: Ethics and governance of artificial intelligence for health, consultabile a questo link, visitato il 9 maggio 2025.

[4] Yoshua Bengio et al., "International AI Safety Report" (DSIT 2025/001, 2025); International Scientific Report on the Safety of Advanced AI: Interim Report, consultabile a questo link, visitato il 9 maggio 2025.

[5] DeepSeek, visitato il 9 maggio 2025.

[6] Xudong Pan, Jiarun Dai, Yihe Fan, and Min Yang. 'Frontier AI Systems Have Surpassed the Self-Replicating Red Line'. arXiv:2412.12140, 9 December 2024, consultabile a questo link.

[7] Leading the Future of Global Health with Responsible Artificial Intelligence'. World Health Organization, 2024, consultabile a questo link.

[8] Giulia Semenzato "Alignment of Reporting Guidelines for Artificial Intelligence in Healthcare with Governance Principles by the World Health Organization: A scoping review". Universita' degli Studi di Parma, Tesi di Master, a.a. 2023/24.

[9] Giulia Semenzato, Mario Ettore Giardini, Cristina Cenci, Carlo Maria Petrini, Luigi Rovati "Aderenza delle linee guida di reporting sull'IA in Sanità ai principi di governance OMS" Congresso Nazionale Società Italiana di Telemedicina – SIT 2025, Bologna 29-31 maggio 2025.

[10] GOV.UK. 'The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023'. Visitato il 24 aprile 2025.