LUISS

# Intelligenza artificiale e diagnostica medica. Tra nuove sfide ed “effetto farfalla”

**di Giovanni Alliegro, Fosco Gentili,**
**Giuseppe Modugno, Alessandro Rondinone**
Luiss School of Government

**Policy Brief n. 16/2022**

*L'avvento dei sistemi intelligenti nel settore sanitario lascia spazio a vantaggi significativi, tra cui spiccano l'efficientamento delle risorse e la "personalizzazione" dell'esperienza terapeutica. La travolgente diffusione dell’intelligenza artificiale (IA) nel campo della diagnostica, tuttavia, esporrà imprese e individui ad una variegata gamma di rischi operativi e legali. Il lavoro che segue pone l’attenzione sul fenomeno del “data poisoning” (o “inquinamento dei dati”), un attacco cibernetico assimilabile al battito d’ali della farfalla capace di generare scenari ad alto impatto. Nel rispondere a tale rischio, saranno affrontate alcune contromisure tecniche e organizzative in ambito di information security e safety, per terminare con un’analisi delle implicazioni in termini di ripartizione della responsabilità medica in caso di danno causato dal sistema IA compromesso.*

Nella teoria del caos, l'effetto farfalla descrive il principio secondo cui variazioni minime nelle condizioni iniziali sono in grado di produrre grandi cambiamenti nel comportamento globale di un sistema. Si ritiene, così, che il battito d'ali di una farfalla sia in grado di provocare un uragano dall'altra parte del mondo.

L'utilizzo dell'Intelligenza Artificiale (IA) nell'assistenza sanitaria è destinato a propagarsi rapidamente, in ragione delle molteplici applicazioni: dalla diagnostica per immagini allo screening di tumori, con considerevoli vantaggi di efficientamento delle risorse e dei tempi clinici. Le forze trainanti di questa espansione sono da rinvenirsi nell'incremento di dataset clinici disponibili, nella crescente domanda di percorsi terapeutici personalizzati.

Tuttavia, è importante evidenziare che l'impatto vantaggioso dell'IA è indissolubilmente legato alla capacità di formulare giudizi o previsioni sulla base di un processo di categorizzazione di vastissime quantità di dati.

Gli attacchi di tipo "data poisoning" (o "inquinamento dei dati") sono tra i più rilevanti ostacoli allo sviluppo sicuro della IA: il battito d'ali capace di generare scenari ad alto impatto per aziende e individui.

**Data poisoning e contromisure**

In un attacco *data poisoning* indiretto, l'agente non ha accesso al dataset utilizzato e deve cercare di infettarlo prima della fase di elaborazione. Un attacco diretto determina, invece, l'alterazione del dataset tramite Data Injection o Data Manipulation. Nel Data Injection, gli input malevoli sono inseriti direttamente nel dataset originario, alterando la distribuzione dei dati sottostanti senza modificarne le caratteristiche o le etichette. Nel Data Manipulation, si agisce modificando le etichette (Label Manipulation) o le caratteristiche (Input Manipulation) dei dati.

Le tecniche di difesa specifiche che vengono applicate ai sistemi di Machine Learning sono da intendersi supplementari alle raccomandazioni e best practice di protezione logica e fisica dei sistemi informativi tradizionali (e.g., awareness, IDS/IPS, patching).

Difese contro il data poisoning includono la Data Sanitization, per cui i dati che causano alti tassi di errore nella classificazione vengono rimossi dal training dataset (c.d. "reject on negative impact"), e la Robust Statistics, per cui vengono utilizzati vincoli e tecniche di regolarizzazione al fine di ridurre potenziali distorsioni del modello di apprendimento.

**Data Security & AI Safety: nuovi paradigmi per la compliance**

Tra le numerose considerazioni applicabili agli attacchi di data poisoning è possibile individuare implicazioni di data privacy & security e di AI safety.

Relativamente al primo punto, è opportuno sottolineare l'importanza di predisporre ambienti di test e produzione che garantiscano il rispetto della riservatezza, integrità e disponibilità del dato sul piano tecnico e organizzativo. Inoltre, la necessità di monitorare lo sviluppo sicuro delle IA impone uno scrupoloso controllo dei livelli di data quality.

Sarà fondamentale procedere a regolari esercizi di data mapping. Tale esercizio ha un ruolo essenziale nella determinazione di adeguate finalità e basi giuridiche del trattamento, corretti volumi e periodi di conservazione dei dataset e monitoraggio continuo delle misure di

sicurezza. L'attività di data mapping dovrà comprendere l'analisi puntuale della compliance posture dei destinatari dei dati mediante processi di due diligence e audit.

Dovrà essere garantita di volta in volta la corretta allocazione di ruoli e responsabilità lungo l'intera supply chain, prestando particolare riguardo alla giurisprudenza della Corte di Giustizia europea e delle Autorità di supervisione in materia di incidenti informatici, trasparenza del trattamento e data transfers. L'inottemperanza a tali considerazioni potrebbe porre intere filiere ora sotto la scure del GDPR, ora sotto quello della normativa sul Perimetro Cibernetico Nazionale e, in via speculativa, del Regolamento IA.

Venendo alle considerazioni in materia di safety, il nuovo Regolamento sui Dispositivi Medici stabilisce una classificazione risk-based dei dispositivi medici. I software medicali destinati a fornire supporto a decisioni diagnostico-terapeutiche sono inclusi nella Classe IIa, salvo il caso in cui tali decisioni possano avere un impatto potenzialmente fatale. In tal caso, è applicata la classe di rischio elevato (Classe IIb o III).

È prospettabile che la proliferazione di sistemi di certificazione nell'ambito degli hardware medicali sia affiancata da una maggiore attenzione verso software e firmware. Allo stesso modo, le soluzioni gestite in cloud dovranno offrire un elevato livello di affidabilità e privacy. Pertanto, gli operatori del settore saranno chiamati a confrontarsi con il Cyber Security Act nonché, in prospettiva, con il quadro normativo in materia di sicurezza delle infrastrutture strategiche (NIS 2.0, Perimetro Nazionale Cibernetico).

In chiusura, è opportuno un riferimento alla proposta di Regolamento IA, che include i sistemi biomedici nella categoria delle IA ad "alto rischio", cui segue un apposito processo di marcatura CE. Notabilmente, i produttori dovranno sottoporre le IA biomediche ad un processo di sviluppo e documentazione di misure di governance e sicurezza coerenti e proporzionate al rischio, conducendo di fatto un AI impact assessment.

**Responsabilità medica e IA compromessa: quali conseguenze?**

Posto quanto riportato, è necessario interrogarsi circa le implicazioni penalistiche discendenti da un attacco di data poisoning su IA diagnostiche. L'esempio è quello della morte di un paziente a seguito del pedissequo rispetto da parte del medico delle indicazioni fornite dalla IA corrotta.

Come accennato, il pericolo di una diffusione incontrollata delle IA diagnostiche coincide con un pericolo di progressiva deresponsabilizzazione dell'agente, con forte tensione sui principi di colpevolezza e di affidamento.

Il principio di colpevolezza afferma che un soggetto, affinché possa considerarsi giuridicamente responsabile, deve aver posto in essere una condotta rimproverabile. Quanto al principio di affidamento, invece, ciascun soggetto può e deve poter confidare nel corretto comportamento altrui, non facendo gravare sui singoli l'onere di rispettare le regole cautelari che li riguardano e di controllare l'osservanza dei doveri cautelari degli altri.

Occorre quindi chiedersi se il medico sia scusabile per legittimo affidamento al sistema di IA oppure se debba riconoscersi un ulteriore dovere cautelare di critica e revisione dell'output.

Una prima ricostruzione afferma la possibilità di riconoscere un legittimo affidamento in quanto l'IA ottimizza i propri output mediante un processo di apprendimento condizionato che dovrebbe renderlo estremamente affidabile. Inoltre, argomentando al contrario sembrerebbe porsi in capo al medico l'obbligo di analizzare di volta in volta tutti possibili esiti alternativi, ripercorrendo a ritroso il "ragionamento" dell'algoritmo. Ne conseguirebbe una distensione inaccettabile delle tempistiche cliniche.

Una seconda ricostruzione, avallata dal Parlamento europeo, nega l'ammissibilità di un legittimo affidamento alle IA. Le IA, infatti, non possono sostituire la decisione umana e gli uomini non sono legittimati, in ragione di ciò, a venir meno ai propri obblighi e alle connesse responsabilità. E ancora, la Legge Gelli-Bianco riporta come il rispetto pedissequo delle linee guida non escluda la colpa ove queste siano inadeguate al caso concreto.

Importante, d'altro canto, definire se l'IA possa costituire un autonomo centro di imputazione giuridica, ipotesi contestata dalla dottrina maggioritaria e radicalmente esclusa dall'attuale quadro normativo. L'operato dell'IA, infatti, non si sostanzia in una scelta libera e consapevole, ma nell'esecuzione di un codice di programmazione, ancorché potenzialmente sorvegliato dall'uomo. Manca, dunque, l'elemento soggettivo identificabile come coscienza dell'agente, in assenza della quale non vi è responsabilità.

In conclusione, la sanzione penale nei confronti della IA risulta priva della sua capacità rieducativa, deterrente e preventiva.

**Cenni finali**

Le iniziative di seguito, già in parte discusse dal Consiglio Superiore di Sanità, potrebbero bilanciare la diffusione di soluzioni IA a supporto della diagnostica:

- sviluppo di una Italian Health Data Strategy: un quadro strategico armonizzato agli obiettivi europei in materia di sicurezza delle infrastrutture e data governance per il settore sanitario;
- istituzione di un'agenzia europea di governance dei sistemi di IA che riporti regolarmente alle autorità nazionali competenti (AIFA, Ministero della Salute, ACN) su questioni attinenti allo sviluppo di IA sanitarie;
- agevolazione di iniziative di scambio di record digitali sanitari per la costituzione e mantenimento di open data lakes, presidiando le fonti informative nel rispetto della normativa in materia di privacy, proprietà intellettuale e industriale mediante appositi organismi di controllo;
- predisposizione di linee guida per il processo decisionale diagnostico-terapeutico del personale medico ove assistito da sistemi di IA;

A discapito dell'eterogeneità del quadro normativo attuale, gli interessi presidiati dal legislatore risultano convergenti. Tuttavia, nella tumultuosa diffusione delle IA, sarà fondamentale indirizzare la "forza propulsiva" del progresso verso soluzioni ispirate a principi virtuosi. Se l'effetto farfalla, più che un'immagine del futuro, impone un "ritorno al passato", l'adozione di un approccio interpretativo trasversale e integrato alle IA è lo strumento essenziale non tanto a domare l'uragano, quanto a prevenire il battito d'ali.