POLICY PAPER

Buone pratiche e modelli di regolamentazione per l'impiego della IA nella diagnostica per immagini

STRUMENTI COMUNI

Definizione:

In generale, con il termine IA ci si riferisce a sistemi con i quali si cerca di imitare l'essere umano e le sue capacità cognitive, ivi compresa la capacità di apprendimento. La maggior parte dei sistemi di IA si basano su approcci di apprendimento automatico supervisionato e non supervisionato. Tali sistemi hanno la capacità di apprendere dei *pattern* nei dati e risolvere particolari *task* a supporto dell'umano che può sfruttare la conoscenza estratta dai dati per assumere proprie decisioni.

Una definizione puntuale è quella di Marco Somalvico (uno dei padri italiani dell'IA), secondo cui l'IA è «*una disciplina informatica che studia i fondamenti teorici, le metodologie e le tecniche che consentono la progettazione di sistemi hardware/software in grado di fornire all'elaboratore elettronico prestazioni che, ad un osservatore comune, sembrerebbero essere di pertinenza esclusiva dell'intelligenza umana*».

Per il diritto dell'Unione Europea, si definisce "sistema di intelligenza artificiale": «*Un sistema automatizzato progettato per funzionare con livelli di autonomia variabili e che può presentare adattabilità dopo la diffusione e che, per obiettivi espliciti o impliciti, deduce dall'input che riceve come generare output quali previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali*» (cfr. art. 3, co. 1, n. 1, AI Act).

Più specificatamente, nel settore sanitario, il termine IA è utilizzato per indicare dispositivi o sistemi in grado di percepire ed elaborare informazioni relative all'ambiente in cui operano e utilizzarle per raggiungere un obiettivo (*task*) predefinito.

Fonti normative:

- Regolamento del Parlamento europeo e del Consiglio sull'intelligenza artificiale del 13.6.2024 (c.d. AI Act) - https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=OJ%3AL_202401689
- Sito internet ufficiale dell'AI Act - https://artificialintelligenceact.eu/
- Proposta di Direttiva del Parlamento europeo e del Consiglio sulla responsabilità da intelligenza artificiale del 28.9.2022 - https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:52022PC0496

- Direttiva del Parlamento europeo e del Consiglio sulla responsabilità per danno da prodotti difettosi del 23.10.2024 - https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=OJ:L_202402853
- Libro bianco sull'intelligenza artificiale del 19.2.2020 - https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:52020DC0065
- Artificial Intelligence in Healthcare. Applications, risks, and ethical and social implications. European Parliament, June 2022 - https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729512/EPRS_STU(2022)729512_EN.pdf
- Disegno di legge in materia di intelligenza artificiale del 20.5.2024 - https://www.senato.it/service/PDF/PDFServer/BGT/01418921.pdf
- Regolamento (UE) 2017/745 del Parlamento europeo e del Consiglio, del 5.6.2017, relativo ai dispositivi medici - https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32017R0745

AMBITO GIURIDICO

Profili di diritto penale

NOZIONI INTRODUTTIVE:

La rilevanza penalistica dell'impiego delle IA in medicina dipende, anzitutto, dal tipo di relazione intercorrente tra IA, paziente e medico, dovendosi distinguere i casi di: 1) rapporto diretto tra IA e paziente, senza l'intermediazione del medico (**modello della IA "indipendente"**); 2) rapporto diretto tra paziente e medico, il quale è supportato dalla IA (**modello della IA "ausiliaria"**). Mentre il primo modello appare ancora meramente ipotetico, il secondo risulta già operativo in numerosi contesti sanitari, a seconda dei quali si potrà ulteriormente distinguere tra: a) **IA di "ausilio tecnico"**; b) **IA di "ausilio intellettuale"**.

I sistemi di IA "di ausilio tecnico" sono equiparabili a qualsiasi attrezzatura sanitaria a disposizione del medico con lo scopo di migliorarne le capacità esecutive: a questi fini, un comune bisturi potrebbe essere equiparato ad un braccio meccanico implementato con tecnologie di IA. Per rimanere nell'ambito della diagnostica per immagini, si pensi a sistemi di analisi MRI/CT "*AI-driven*", ove l'intelligenza artificiale interviene a migliorare la qualità delle immagini da impiegare per la valutazione clinica del paziente. In simili casi, la suddivisione della responsabilità penale tra produttore di IA di "ausilio tecnico" e medico è lineare: se il medico erra nell'uso della IA, risponderà il medico; se il medico erra a causa di un malfunzionamento della IA, risponderà il produttore ove si accerti un difetto di progettazione.

I sistemi di IA "di ausilio intellettuale" supportano il processo decisionale del medico, proponendo anche in via del tutto autonoma delle possibili soluzioni cliniche: si pensi per esempio all'utilizzo di reti neurali convoluzionali (CNN) per la classificazione delle immagini, o ancora all'impiego di tecniche di *deep learning* per implementare l'analisi di immagini radiologiche. Questi sistemi non si limitano a migliorare la raccolta dei dati clinici, ma sono addirittura in grado di svolgere attività di tipo compiutamente "cognitivo". Pertanto, in simili ipotesi il medico non viene meramente chiamato soltanto a "utilizzare" la IA (prestando legittimo affidamento sulle capacità tecniche della stessa), bensì pure a "confrontarsi" con i suoi output valutativi. La suddivisione della responsabilità penale tra produttore di IA di "ausilio intellettuale" e medico non è lineare: se il medico erra nell'uso della IA o si discosta erroneamente dalla soluzione proposta dalla IA, risponderà il medico; se il medico aderisce erroneamente alla soluzione errata proposta dalla IA, risponderà il medico, al netto di eventuali responsabilità del produttore in caso di difetti di progettazione.

Sono questi ultimi i sistemi che pongono maggiori problemi dal punto di vista della responsabilità.

PROBLEMATICHE APPLICATIVE:

Con riguardo alla responsabilità penale del medico radiologo, le specifiche problematiche applicative derivanti dall'impiego dei sistemi di IA "di ausilio

intellettuale" nella diagnostica per immagini dipendono in larga parte dal fatto che **se, da un lato, le diagnosi delle IA possono risultare anche statisticamente più accurate rispetto a quelle umane, dall'altro, tuttavia, attualmente non è ancora possibile risalire al "ragionamento" della macchina**, posto tra l'altro che essa si basa su inferenze non causali, e si limita a comunicare all'utenza dei risultati "secchi", cioè privi di motivazione.

Tali caratteristiche appaiono idonee a incidere sul **giudizio controfattuale** e sulla correlata **"esigibilità" della condotta alternativa lecita**, su cui si fonda la responsabilità colposa, con conseguente necessità di interrogarsi sui seguenti quesiti: **1)** se il medico e la IA giungono ad una stessa diagnosi, la quale, però, si rivela errata, può considerarsi "esigibile" la diagnosi corretta (quando persino la IA non è riuscita ad elaborarla)? **2)** Se il medico utilizza erroneamente il sistema di IA, il medico deve essere sempre considerato responsabile per l'errore diagnostico? **3)** Se il medico si discosta dalla soluzione offerta dalla IA, la quale, però, si rivela corretta, il medico deve essere sempre considerato responsabile per l'errore diagnostico?

Nessun caso può essere risolto tramite automatismi: **1)** la conferma della diagnosi da parte della IA può rappresentare un elemento di prova dell'inesigibilità della diagnosi corretta, ma in caso di errori macroscopici della IA non sarebbe ipotizzabile un effetto esimente; **2)** l'erroneo impiego della IA da parte del medico può fondare la responsabilità colposa, ma è necessario dimostrare che il corretto utilizzo della IA avrebbe evitato l'errore diagnostico; **3)** il dissenso rispetto alla soluzione proposta dalla IA può fondare la responsabilità colposa, ma è necessario dimostrare che le motivazioni del dissenso non fossero ragionevoli.

Tutto ciò, fermo restando che un errore fondamentalmente "rimproverabile", quindi non tale da escludere la responsabilità (e però almeno in parte "scusabile" perché indotto dall'interazione con IA), potrebbe comunque spingere il giudice ad una quantificazione della pena verso i minimi edittali, ai sensi dell'art. 133 c.p., o all'applicazione delle "attenuanti generiche" di cui all'art. 62-*bis* c.p.

Dal punto di vista politico-criminale, la natura delle inferenze dei sistemi IA (non vincolate dalle leggi scientifiche mediche) e dei suoi esiti (l'*output* è immediatamente correlato al caso clinico che si sta affrontando e non abbisogna di alcuna concretizzazione) pone il problema della perdurante attualità e razionalità del regime della colpa medica previsto dalla l. n. 24 del 2017. D'altra parte, non sarebbe auspicabile che l'impiego delle IA nella diagnostica per immagini sia regolato con meri automatismi giuridici: se la conferma delle IA rappresentasse sempre un'esimente per il medico, quest'ultimo finirebbe per appiattirsi sulle valutazioni delle IA, rinunciando alla propria autonomia valutativa; d'altra parte, se al minimo errore esecutivo o dissenso rispetto alle soluzioni delle IA conseguisse sempre un addebito penale per il medico, quest'ultimo finirebbe per preferire contesti sanitari privi di sistemi di IA.

Per assicurare la corretta operatività del diritto penale, **è necessaria l'implementazione di regole di utilizzo clinico uniformi, e di modelli di sviluppo informatico adeguati all'utilizzo in ambito clinico**: dal punto di vista dell'organizzazione sanitaria, le procedure diagnostiche, l'impiego tecnico-

operativo delle IA e le modalità di espressione del dissenso non dovrebbero essere affidate alla sensibilità dei singoli medici, bensì codificate secondo logiche di adeguatezza clinica e d'interesse collettivo; dal punto di vista informatico, è fondamentale che i dispositivi medici IA siano il più possibile armonici con i principi fondamentali che caratterizzano il peculiare contesto assiologico in cui si inseriscono, e dunque siano in grado, tra l'altro, di valorizzare l'autodeterminazione terapeutica del paziente e altresì la professionalità clinica del medico (ad esempio, rendendo più trasparenti le ragioni della decisione, garantendo una certa interattività del suo *output*, anche per non lasciare privo di "tutela sanitaria digitale" il paziente che non aderisce alla prima "miglior soluzione obiettiva" proposta dalla macchina, etc.); inoltre, è necessario disporre di mezzi spendibili a livello processuale per dimostrare se l'errore della macchina sia dovuto a un malfunzionamento o all'obiettiva difficoltà del caso, nonché per dimostrare, in ottica controfattuale, che il corretto utilizzo della IA – secondo il livello di sviluppo tecnologico al momento dei fatti – avrebbe evitato l'errore diagnostico.

Con riguardo alla responsabilità del produttore della IA, sussistono alcuni ostacoli teorici a che l'errore diagnostico della medesima possa essergli penalmente attribuito (anche a prescindere dalle difficoltà pratiche rispetto all'accertamento del nesso di causalità): da un lato, l'elaborazione dell'immagine da parte della IA, non essendo equiparabile alla diagnosi di un radiologo, potrebbe difettare della tipicità causale necessaria per integrare un'ipotesi di **concorso di cause**; dall'altro lato, parrebbe complesso ipotizzare una forma di **cooperazione colposa**, sul modello del lavoro di *équipe*, poiché tale istituto presuppone la volontà di concorrere alla realizzazione della condotta contraria alle regole cautelari, elemento che potrebbe non considerarsi ravvisabile in capo al produttore della IA.

Qualora si superassero tali ostacoli teorici, bisognerebbe poi accertare il **nesso di causalità tra l'eventuale difetto di produzione e l'errore diagnostico della IA**; tale accertamento è complicato dal fatto che le diagnosi "secche" formulate dalle IA non consentono di individuare in modo agevole le cause (malfunzionamento della IA o obiettiva complessità della diagnosi) di un eventuale errore diagnostico, sempre che quest'ultimo, ad esempio, non ricorra in modo sistematico.

Infine, qualora si fosse **in presenza di un sistema di IA in *continual learning***, una volta individuato l'eventuale malfunzionamento, bisognerebbe anche dimostrare che quest'ultimo dipenda da un effettivo difetto progettuale e non, invece, alla fisiologia del processo di *self learning*.

Profili di diritto privato

NOZIONI INTRODUTTIVE:

La progressiva diffusione dell'intelligenza artificiale in ambito medico impone al civilista di sondare la capacità di tenuta delle "classiche" categorie di diritto privato di fronte ai nuovi interrogativi posti dall'innovazione tecnologica. Anche per quanto concerne l'ambito privatistico, all'attuale stato dell'evoluzione tecnico-

scientifica, a destare il maggiore interesse sono senza dubbio **i sistemi di IA di "ausilio intellettuale".**

Ogni indagine giuridica che si rispetti deve necessariamente partire da un'opera di puntuale classificazione del proprio oggetto di studio. Al fine di comprendere al meglio la corretta classificazione dei sistemi di IA di "ausilio intellettuale" utilizzati nella diagnostica per immagini è utile immaginare di avere di fronte a sé una *matrioska*: in prima battuta, tali sistemi devono essere ricondotti alla categoria dei beni mobili; poi, all'interno di questa prima macrocategoria, gli stessi rientrano nella nozione di prodotti; ulteriormente, vista la loro concreta utilizzazione, tali sistemi afferiscono al successivo sottoinsieme dei dispositivi medici; infine, all'interno dei dispositivi medici, quelli incorporanti sistemi di intelligenza devono essere classificati sotto l'etichetta di dispositivi medici ad alto rischio.

Parlando ancora in termini generali, occorre altresì ricordare come, a differenza della responsabilità penale, quella civile possa prescindere dalla valutazione e dalla prova del dolo o della colpa del soggetto agente (danneggiante). Questa significativa circostanza comporta un'assai maggiore libertà ricostruttiva da parte dell'interprete. Inoltre, sempre in tema di responsabilità *lato sensu* civile, è importante tener presente che, nell'evenienza di danni sanitari, a venire potenzialmente in gioco sono tanto la responsabilità contrattuale della struttura sanitaria, quanto quella extracontrattuale del medico (con necessità di provare il dolo o la colpa del sanitario).

In tema di responsabilità, è poi necessario mettere in evidenza come, in caso di esito infausto del trattamento sanitario, qualora questo sia stato determinato in tutto o in parte da un malfunzionamento del dispositivo di IA, potranno essere considerati responsabili del danno anche i soggetti afferenti alla composita attività di produzione del prodotto.

PROBLEMATICHE APPLICATIVE:

Al di là delle difficoltà derivanti dall'esistenza di un quadro normativo assai composito e in rapido cambiamento, molte delle problematiche legate all'utilizzo dei sistemi di IA in ambito diagnostico ruotano intorno al grande tema della **"responsabilità".**

Ai fini di una maggiore chiarezza espositiva, i cangianti profili di responsabilità in gioco possono essere schematicamente suddivisi in due direttrici principali: da un lato, la responsabilità dei vari soggetti che hanno partecipato alla fase di produzione; dall'altro, quella della struttura sanitaria e del medico.

Concentrandosi, dapprima, **sulla responsabilità afferente alla fase di produzione**, è stato da molti messo in luce come, in caso di danno per malfunzionamento, data la complessità del prodotto e la difficile intellegibilità dei suoi meccanismi operativi e di autoapprendimento, per il danneggiato risulti estremamente arduo, anche alla luce della grande asimmetria informativa: dimostrare il difetto del prodotto, comprendere l'eziologia del malfunzionamento e individuare quale fra i vari soggetti che hanno contribuito allo sviluppo del sistema di IA debba essere ritenuto responsabile. Infatti, in merito a quest'ultimo

aspetto, è stato evidenziato come, oltre al formale produttore del *device* e delle sue varie componenti fisiche, nel processo di costruzione e sviluppo del dispositivo intervengano anche: il programmatore/autore dell'algoritmo; l'addestratore della macchina; e, infine, il soggetto che si occupa di fornire i dati necessari per l'addestramento del sistema.

Sul punto, va segnalato che la nuova direttiva sulla responsabilità da prodotto difettoso migliora sensibilmente le condizioni processuali dei danneggiati: stabilendo a loro vantaggio un diritto di *disclosure*; prevedendo significative agevolazioni probatorie; introducendo il concetto di difetto sopravvenuto per i prodotti su cui il fabbricante continua a esercitare il controllo; e limitando l'applicabilità della difesa da rischio da sviluppo. Tuttavia, il complesso funzionamento dell'AI, la necessità di provare il difetto del prodotto, la permanenza (seppur limitata) della difesa da rischio da sviluppo e il macchinoso sistema di presunzioni e contro-presunzioni continueranno a compromettere le reali possibilità di successo processuale dei danneggiati.

Tali difficoltà probatorie sono proprie, in realtà, di molti prodotti altamente tecnologici, specialmente se integranti l'IA. Per questo, alcuni autori suggeriscono un generale ripensamento del criterio di imputazione della responsabilità, ovvero la scelta di individuare a monte – fra i vari soggetti che compartecipano alla produzione – quello meglio in grado di farsi carico del risarcimento, legittimando il danneggiato ad agire direttamente nei suoi confronti e lasciando poi spazio a successivi meccanismi di regresso. Altri consigliano, invece, l'adozione di un particolare tipo di responsabilità vicaria in cui a farsi carico dei danni, indipendentemente dalla prova del difetto o da qualsiasi forma di colpa, siano, almeno in prima istanza, non i produttori ma i soggetti imprenditoriali che si avvalgono e traggono vantaggio dall'utilizzo dell'IA. Applicando quest'ultima ricostruzione ai dispositivi medici per la diagnostica per immagini, in caso di danno al paziente, in prima battuta, il soggetto responsabile dovrebbe sempre essere la struttura sanitaria, la quale, in presenza di evidenti malfunzionamenti, potrà, poi, rivalersi in un secondo momento sul produttore.

Peraltro, anche qualora a nessuno dei soggetti operanti nella fase di produzione possa essere rimproverato alcunché (è il caso in cui l'erroneo responso della macchina non può essere considerato un malfunzionamento e/o un difetto, ma un preventivabile e inevitabile fallimento operativo), dato che nel diritto civile è pacificamente ammessa la c.d. responsabilità oggettiva, alcuni osservatori raccomandano che il produttore del dispositivo sia in ogni caso considerato responsabile del danno. Tuttavia, in ottica di politica del diritto, risulta necessario chiedersi se – visti gli indubbi miglioramenti offerti alla scienza medica e, di riflesso, alla salute dei pazienti dai sistemi di IA basati sul *machine learning* – sia auspicabile che il costo di questi inevitabili danni venga sostenuto dai produttori o se, invece, anche per evitare un possibile *chilling effect* sul settore, non sia preferibile che di tale onere si facciano carico la collettività e/o le strutture sanitarie.

Riflessioni diverse si legano, invece, ai profili di responsabilità del medico. In particolare, molti studiosi sottolineano che l'introduzione dei sistemi di IA di ausilio intellettuale pare in grado di mettere quantomeno in discussione le

fondamenta della relazione medico-paziente e, con essa, il giudizio di colpevolezza sull'operato del sanitario. Difatti, se fino ad oggi il medico e la sua capacità decisionale sono state "semplicemente" coadiuvate da dispositivi di supporto, di converso, i sistemi di IA sono capaci di emettere responsi diagnostici – non sempre di facile comprensione – frutto di un processo di apprendimento dinamico e basato sullo studio di enormi quantità di dati. Pertanto, considerate la mole di informazioni conosciute dal sistema, assai maggiore di quella padroneggiata da un medico esperto, la "plasticità" del dispositivo, ossia la capacità di apprendere generativamente grazie a nuove informazioni, e la non semplice intelligibilità dei processi decisionali della macchina, quando la decisione del medico di discostarsi dal responso offerto dal sistema di IA potrà dirsi giustificata? A fronte di meccanismi di funzionamento e apprendimento del sistema di difficile intellezione, come potrà il medico motivare la propria scelta terapeutica? Quale deve considerarsi lo standard di condotta medica corretto? Molti interpreti ravvisano il pericolo di un medico che non sia più il protagonista della prestazione di cura, bensì un mero supervisore acritico dell'opera del dispositivo. Peraltro, come osservato dalla dottrina, sussiste il concreto rischio che, in caso di esito infausto, i pazienti, quasi di *default*, esercitino contro i sanitari azioni di responsabilità per non avere utilizzato o seguito le indicazioni del sistema di IA. Questo potrebbe portare il medico a aderire acriticamente alle decisioni del dispositivo, determinando *"una nuova frontiera della medicina difensiva"* e la conseguente scomparsa di ogni autonomia del sanitario.

Alla luce di questo, è chiaro come il futuro e le sorti della responsabilità civile del medico ruotino intorno al concetto di trasparenza. Segnatamente, come messo in luce da molti autori, maggiore è la trasparenza del sistema di IA e con essa il controllo esercitabile sullo stesso da parte dell'operatore, più ampi sono i confini della responsabilità del medico. Di converso, minore è la trasparenza, maggiori sono i profili di responsabilità della struttura sanitaria e del produttore. In ragione di ciò, si potrebbe persino sostenere che, ogniqualvolta la struttura opti per l'adozione di sistemi non sufficientemente trasparenti, essa debba sempre essere considerata l'unica responsabile per eventuali diagnosi errate. Tuttavia, la trasparenza del processo decisionale del sistema, l'effettivo controllo e la responsabilità del sanitario sembrano fattori indispensabili al fine di preservare le basi dell'importantissima relazione medico-paziente. Difatti, qualora la mancanza di trasparenza esautorasse – anche solo in apparenza – il medico del suo effettivo ruolo decisionale e della responsabilità delle scelte da lui prese, questo potrebbe minare le fondamenta del rapporto di fiducia col paziente. All'opposto, è stato ben evidenziato che – se realmente trasparenti (con ciò intendendosi comprensibilità del funzionamento e intellegibilità del processo decisionale e di ragionamento) – tali dispositivi potrebbero essere un veicolo per rafforzare la relazione di cura. Infatti, essi possono sollevare i medici da numerosi compiti, permettendo loro di dedicare più tempo al dialogo con i pazienti.

Concludendo sul tema, va segnalato che, sulla scia del noto scandalo PIP, un ultimo profilo di responsabilità potrebbe essere quello degli organismi notificati. Infatti, i dispositivi medici – in particolare quelli ad alto rischio – sono soggetti a stringenti valutazioni di sicurezza e conformità da parte di specifici *notified bodies*. Pertanto, qualora tali enti errino nella loro attività di vigilanza e in conseguenza

di ciò si verifichi un danno, essi potrebbero essere ritenuti responsabili sulla base di ipotesi di responsabilità civile previste dagli ordinamenti nazionali.

Un altro profilo su cui i sistemi di IA impattano sensibilmente è quello del **consenso informato al trattamento medico**. Infatti, nel corso degli ultimi anni, quello che era nato come un nobile strumento a tutela del paziente si è pian piano tramutato in un vero e proprio dogma da rispettare al fine di mettersi al riparo da ogni responsabilità. In virtù di questo e vista la difficoltà di comprendere il processo decisionale sotteso al responso della macchina, è necessario chiedersi se e come il paziente debba essere informato del ruolo ricoperto dal sistema e del modo in cui la decisione finale viene presa. In particolare, la civilistica cerca di rispondere alle seguenti domande: il paziente dovrà in ogni caso essere informato dell'utilizzo dell'intelligenza artificiale? Dovrà conoscere il grado di trasparenza sotteso al responso del dispositivo? Dovrà essere edotto del funzionamento dello stesso? Dovrà essere informato del fatto che, in una determinata percentuale di casi, il sistema formula responsi errati? Spetterà al paziente o al medico scegliere se discostarsi dal responso del dispositivo? Spingendosi ancora oltre, nel caso in cui una struttura utilizzi abitualmente l'IA in determinati protocolli terapeutici, il paziente avrà il diritto di rifiutare tale trattamento e di pretenderne uno alternativo, privo dell'impiego di tale tecnologia?

Infine, **in tema di consenso al trattamento dei dati sanitari**, alcuni studiosi hanno osservato come il funzionamento del *machine learning* imponga – alla luce della solidarietà e del diritto alla salute collettiva – una riflessione sui limiti alla possibilità di negare il consenso da parte del paziente. Infatti, visto che il sistema per poter apprendere e diventare sempre più "bravo" e preciso ha bisogno di un innumerevole quantità di dati, è accettabile che i pazienti che usufruiscono dell'opera dello stesso si rifiutino, poi, di prestare il consenso al trattamento dei propri dati per fini sanitari e di ricerca? In altri termini, è giusto che chi trae vantaggio da un sistema basato sulle informazioni raccolte grazie ad altri pazienti possa legittimamente negare il proprio consenso, così privando i futuri pazienti (e la collettività) della possibilità di avere diagnosi sempre più accurate?

Quesiti

DA PARTE DEI MEDICI:

1) *Che tipo di effetto potrebbe avere l'introduzione dell'IA sulle normative che regolano l'errore medico? Supponiamo il caso di una mancata diagnosi: se l'errore del medico fosse in accordo con il software (ovvero, anche l'IA ha sbagliato la diagnosi) potrebbe rappresentare un'attenuante o un'aggravante? E se invece l'errore fosse in disaccordo con il software (ovvero l'IA ha formulato correttamente la diagnosi, ma il medico ha deciso di non seguire l'output dell'algoritmo)? Se così fosse, l'IA potrebbe favorire un'ulteriore diffusione dell'ormai, purtroppo, consolidata medicina difensiva. In tal senso, tocca al medico farsi carico del rischio di questo confronto, o il legislatore può in qualche modo prevedere questa evenienza?*

Preliminarmente, occorre precisare che, in diritto penale, il concetto di attenuante o aggravante si riferisce a specifiche circostanze accessorie rispetto al

reato, da cui, per espressa previsione legislativa, discendono aumenti o diminuzioni sul piano sanzionatorio; pertanto, in mancanza di specificazioni legislative in materia di IA e colpa medica, i profili di convergenza o divergenza tra diagnosi umane e artificiali spiegherebbero i loro effetti principalmente in relazione al concreto accertamento della responsabilità del medico per l'errore.

Ciò posto, con riguardo ad una convergenza tra medico e IA verso una diagnosi errata, c.d. diagnosi "errata-conforme", occorre distinguere tra situazioni ordinarie, in cui è possibile ravvisare in capo al medico un legittimo affidamento esclusivamente sul funzionamento del macchinario, ed altre, più rare, se non del tutto eccezionali (perché non corrispondenti alla metodologia diagnostica più accreditata), in cui tale affidamento debba estendersi anche alla validità clinica dell'esito diagnostico della IA; ciò potrebbe accadere quando, ad esempio, la valutazione del professionista non possa competere con il grado di accuratezza di quella formulata dalla IA, oppure quando l'IA realizzi inferenze al di fuori della portata del singolo professionista (sistemi di diagnosi IA in grado di rilevare un tumore al senso dalle sole immagini radiologiche e senza bisogno di biopsia, etc.).

Nel primo caso, la presenza di un esito omologo del sistema di IA potrà essere presa in considerazione ai fini di un'attenuazione della gravità della colpa, con conseguenti effetti anche sulla commisurazione della pena; nel secondo caso, invece, sarebbe ipotizzabile un esonero da responsabilità del medico che abbia prestato incolpevole affidamento sulla diagnosi della IA.

A tale ultimo proposito, lo schema potrebbe essere assimilabile a quanto la giurisprudenza già fa con riferimento alla responsabilità penale colposa del medico che opera in équipe (dunque in contesti ove è richiesta l'azione sinergica di più persone e più ruoli): da tempo si afferma, infatti, che l'affidamento ragionevolmente riposto dal professionista nella corretta esecuzione delle proprie mansioni da parte dei colleghi esclude la sua colpa quando costoro, invece, col loro sbaglio, lo abbiano indotto in errore.

Con riguardo, invece, all'ipotesi della diagnosi "errata-divergente", occorre anzitutto evidenziare che, al momento della formulazione del dissenso rispetto alla diagnosi della IA, risulta cruciale la presenza di una motivazione basata su elementi specifici e circostanziati, debitamente evidenziata nella documentazione clinica (così da poterla poi meglio spendere in una eventuale sede processuale). Invero, un generico o comunque immotivato dissenso, tanto più se alimentato da una acritica sfiducia verso la tecnologia in questione, dovrebbe essere considerato di per sé una negligenza fonte di addebito in caso di evento infausto.

Ciò detto, il caso della decisione errata-divergente rispetto a un output diagnostico corretto della macchina pone il clinico in una situazione critica: in casi del genere, infatti, il giudice ha immediatamente a disposizione la "controprova" della possibilità di un risultato diagnostico alternativo e fausto per il paziente, rendendo perciò più complesso un esonero di responsabilità per il medico.

Pertanto, per evitare la formazione di pratiche di "medicina difensiva", risulterebbe cruciale una chiara regolamentazione delle modalità di emersione del dissenso da parte del medico, che lo possano tutelare quantomeno in presenza di diagnosi errate-divergenti "inevitabili", cioè che sarebbero state condivise da qualunque altro medico, pur prestando la diligenza esigibile in quelle stesse condizioni.

In prospettiva di riforma, si potrebbe altresì ragionare su di una regolamentazione che prenda a modello la legge c.d. "Gelli-Bianco" ed il correlato art. 590-*sexies* c.p.: come noto, talune letture la definiscono un'esimente *ex lege* laddove il medico si sia "affidato" a linee guida accreditate da un Sistema nazionale, salvo che "le specificità del caso concreto" avessero dovuto indurlo a discostarsene. Ecco, oggi una simile ipotesi, tutta da indagare, dovrebbe però problematicamente tener conto dell'esito diagnostico restituito dai sistemi di IA; un esito che, se non fosse assimilabile alle raccomandazioni di linee guida, potrebbe tuttavia esser fondatamente valorizzato, a certe e ben definite condizioni, nel giudizio di responsabilità del sanitario. Del resto, il sistema allestito dalla legge "Gelli-Bianco" ha dato esiti largamente insoddisfacenti, tanto sul piano penale che civile, e potrebbe perciò beneficiare di una riforma intesa a meglio valorizzare anche le istanze della c.d. medicina "guidata dai dati", destinata a diventare protagonista nel prossimo futuro.

2) *Il corretto funzionamento di sistemi di IA si basa sulla raccolta ed elaborazione di big data, sia relativi alle immagini biomediche che a dati "non imaging" (p.es. anamnesi, obiettività clinica, parametri laboratoristici, istologici e biomolecolari, informazioni prognostiche), ottenuti da un grande numero di pazienti e la cui condivisione in forma pseudonimizzata/anonimizzata tra più centri di ricerca rappresenta una condizione indispensabile per la conduzione di trials multicentrici, di vitale importanza per il progresso medico-scientifico e il miglioramento della sanità pubblica e individuale. Vi sono spazi di compatibilità tra questo scenario più multidisciplinare e "universale" da un lato e, dall'altro, la legislazione sulla protezione dei dati personali e, in particolare, le attuali modalità di acquisizione del consenso informato al trattamento dei dati?*

Per rispondere alla domanda è necessario realizzare una serie di importanti premesse.

In primo luogo, va evidenziato che i sistemi di IA per la diagnostica per immagini raccolgono ed elaborano dati – in particolare dati sanitari – il cui trattamento è regolato, al contempo, dalla normativa di matrice europea e dal diritto italiano.

Il settore è attualmente in profonda evoluzione. **A livello europeo**, nella primavera del 2024 il Parlamento e il Consiglio hanno trovato un accordo provvisorio sul testo della Proposta di Regolamento UE sullo spazio europeo dei dati sanitari (European Health Data Space, d'ora in poi EHDS). **Sul piano interno**, con legge di conversione del 29 aprile 2024, n. 56, il Parlamento italiano ha convertito il decreto-legge 2 marzo 2024, n. 19, con il quale il Governo è intervenuto, tra l'altro, sull'art. 110 del Codice in materia di protezione dei dati personali (da ora in poi, Codice). Inoltre, è attualmente in discussione in Parlamento il disegno di legge del Governo del 23 aprile 2024 (Norme in materia di intelligenza artificiale), il quale, in caso di approvazione, troverà applicazione in materia.

In secondo luogo, occorre distinguere tra **dati personali**, alla cui disciplina sono ricondotti **i dati pseudonimizzati**, e **dati non personali**, nei quali rientrano anche **i dati anonimizzati**. La stringente normativa del General Data Protection Regulation (d'ora in poi GDPR) trova, infatti, applicazione solo in riferimento al **dato personale**, per tale intendendosi «*qualsiasi informazione riguardante una*

persona fisica identificata o identificabile ("interessato"); si considera identificabile la persona fisica che può essere identificata, direttamente o indirettamente, con particolare riferimento a un identificativo come il nome, un numero di identificazione, dati relativi all'ubicazione, un identificativo online o a uno o più elementi caratteristici della sua identità fisica, fisiologica, genetica, psichica, economica, culturale o sociale» (art. 4, n. 1 GDPR).

Sempre ai sensi del GDPR, i dati pseudonimizzati sono, invece, dati personali sottoposti a un procedimento di pseudonimizzazione, trattati cioè in modo da non poter più essere «*attribuiti a un interessato specifico senza l'utilizzo di informazioni aggiuntive, a condizione che tali informazioni aggiuntive siano conservate separatamente e soggette a misure tecniche e organizzative intese a garantire che tali dati personali non siano attribuiti a una persona fisica identificata o identificabile*» (art. 4, n. 5 GDPR).

Dunque, è di fondamentale importanza comprendere quale sia **il confine fra un dato anonimo e un dato pseudonimizzato, essendo soltanto il secondo sottoposto alle regole di trattamento del GDPR.** In merito, per stabilire l'identificabilità di una persona «*è opportuno considerare tutti i mezzi, come l'individuazione, di cui il titolare del trattamento o un terzo può ragionevolmente avvalersi per identificare detta persona fisica direttamente o indirettamente. Per accertare la ragionevole probabilità di utilizzo dei mezzi per identificare la persona fisica, si dovrebbe prendere in considerazione l'insieme dei fattori obiettivi, tra cui i costi e il tempo necessario per l'identificazione, tenendo conto sia delle tecnologie disponibili al momento del trattamento, sia degli sviluppi tecnologici. I principi di protezione dei dati non dovrebbero pertanto applicarsi a informazioni anonime, vale a dire informazioni che non si riferiscono a una persona fisica identificata o identificabile o a dati personali resi sufficientemente anonimi da impedire o da non consentire più l'identificazione dell'interessato*» (Considerando n. 26 GDPR).

Secondo il Garante per la protezione dei dati personali italiano (GPDP), il dato anonimizzato «*è tale solo se non consente in alcun modo l'identificazione diretta o indiretta di una persona, tenuto conto di tutti i mezzi (economici, informazioni, risorse tecnologiche, competenze, tempo) nella disponibilità di chi (titolare o altro soggetto) provi a utilizzare tali strumenti per identificare un interessato*» e un processo di anonimizzazione «*non può definirsi effettivamente tale qualora non risulti idoneo ad impedire che chiunque utilizzi tali dati, in combinazione con i mezzi "ragionevolmente disponibili", possa: 1. isolare una persona in un gruppo (single-out); 2. collegare un dato anonimizzato a dati riferibili a una persona presenti in un distinto insieme di dati (linkability); 3. dedurre nuove informazioni riferibili a una persona da un dato anonimizzato (inference) (cfr. Parere 05/2014 - WP 216 sulle tecniche di anonimizzazione, adottato il 10 aprile 2014)*» (*ex multis*, Provvedimento 7 marzo 2024, n. 10009033).

È inoltre significativo sottolineare che, qualora il titolare del trattamento intenda condividere i dati anonimizzati con la comunità scientifica, come nel caso in esame, il Garante ritiene necessario rafforzare le tecniche di anonimizzazione «*o attraverso l'introduzione di elementi di distorsione del dato (offset) su tutti gli attributi del record, qualora si intenda mantenere singolarità all'interno del dataset, ovvero attraverso l'applicazione della richiamata tecnica di K-anonimity*» (Provvedimento del 26 ottobre 2023, n. 9960973).

Sebbene non di carattere generale, essendo stati adottati in risposta a istanze di consultazione preventiva *ex* art. 110 del Codice, i citati provvedimenti sono in ogni caso utili per comprendere come il Garante italiano interpreta l'anonimizzazione del dato e quali misure tecniche sono richieste affinché il dato possa essere considerato anonimo.

Va ulteriormente sottolineato che, qualora in un unico insieme di dati siano presenti sia dati personali che dati non personali "indissolubilmente legati" fra loro, si applica la disciplina del GDPR all'intero *dataset* (art. 2, par. 2 del Regolamento (UE) 2018/1807 relativo a un quadro applicabile alla libera circolazione dei dati non personali nell'Unione europea). Come indicato dalla Commissione Europea nel 2019 nella Guida al Regolamento (UE) 2018/1807, il legame indissolubile può aversi laddove separare i dati personali da quelli non personali sia impossibile, economicamente inefficiente o non tecnicamente fattibile per il titolare del trattamento, oppure qualora la separazione comporti una diminuzione del valore dell'intero *dataset.*

Anche la proposta di Regolamento sull'EHDS distingue tra dati sanitari elettronici **personali** («*data concerning health and genetic data as defined in Article 4, points (13) and (15), of Regulation (EU) 2016/679, processed in an electronic form*», art. 2, par. 2, lett. a) e dati sanitari elettronici **non personali** («*electronic health data other than personal electronic health data, encompassing both data that has been anonymised so that it no longer relates to an identified or identifiable natural person and data that has never related to a data subject*»**,** art. 2, par. 2, lett. b). Tuttavia, occorre segnalare che il Capo IV contiene specifiche norme che regolano anche i dati sanitari elettronici non personali (v. l'art. 41, par. 6, EHDS che prevede l'obbligo in capo ai titolari dei dati sanitari elettronici non personali di garantirne l'accesso illimitato a tutti gli utenti, nonché l'archiviazione e la conservazione tramite database aperti e affidabili). Pertanto, in caso di futura applicazione, i titolari del trattamento risulteranno gravati di alcuni obblighi di condotta anche in riferimento ai dati sanitari elettronici non personali.

Chiariti tali aspetti, è ora possibile entrare nel merito della domanda. Sul punto, è importante notare fin da subito che la risposta al quesito varia notevolmente in funzione della tipologia di dato oggetto di trattamento.

a) Come visto, **il dato anonimo non è un dato personale.** Pertanto, **esso non rientra nell'alveo applicativo del GDPR** (v. Considerando n. 26) e **non è necessaria alcuna base giuridica per il relativo trattamento** (perciò neanche il consenso)**.** In tale evenienza, lo scenario multidisciplinare e "universale" evocato dalla domanda non è quindi ostacolato dalla disciplina giuridica. Tuttavia, come accennato, occorre considerare che il **dato anonimo può, in alcune circostanze, essere de-anonimizzato. In tal caso, il dato non è più considerato anonimo e il suo trattamento è soggetto alle regole dettate dal GDPR.**

b) Nell'ipotesi in cui, invece, il dato sia pseudonimizzato si applica il GDPR. Nello specifico, la domanda fa riferimento all'attività di condivisione di dati, la quale è una forma di trattamento *ex* art. 4, par. 1, n. 2 GDPR). **Per questo motivo, affinché la condivisione possa essere realizzata lecitamente deve sussistere una base giuridica ai sensi dell'art. 6 GDPR. Inoltre, i dati sanitari sono considerati una categoria particolare di dati personali, per il cui**

trattamento è altresì necessaria la presenza di una delle "eccezioni" elencate nel paragrafo 2 dell'art. 9 GDPR.

Relativamente alle **basi legittime di trattamento individuate dall'art. 6, par. 1, GDPR**, esse sono tra loro alternative e, **in riferimento alla diagnostica per immagini nel settore pubblico**, possono venire in gioco, soprattutto, **il consenso dell'interessato** (art. 6, par. 1, lett. a, GDPR) e **l'interesse pubblico** (art. 6, par. 1, lett. e, GDPR).

Tuttavia, come detto, considerato che il dispositivo medico oggetto di studio tratta prevalentemente dati sanitari, va specificato che, ai fini della liceità del trattamento, il GDPR richiede (oltre alla base giuridica di cui all'art. 6, par.1) anche l'individuazione di una delle **"eccezioni" elencate all'art. 9, par. 2**. In particolare, per poter utilizzare i dati per fini di ricerca occorre: a) che il consenso dell'interessato sia esplicito (art. 9, par. 2, lett. a, GDPR); b) o che il trattamento sia necessario per esigenze di interesse generale (quali la sanità pubblica o la ricerca scientifica) e sia previsto da una disposizione di diritto dell'UE o interno (art. 9, par. 2, lett. g, i, j, GDPR).

In altre parole, considerato che la domanda fa riferimento a un trattamento per fini di ricerca scientifica o per esigenze sanitarie, il titolare del trattamento (nel caso di specie l'università o l'istituto di ricerca) dovrà alternativamente: dimostrare di aver ottenuto un esplicito e specifico consenso del paziente al progetto di ricerca; ovvero individuare una precisa disposizione che autorizza il trattamento.

In concreto, volendo bypassare il consenso dell'interessato, l'individuazione di una disposizione di tal sorta è tutt'altro che agevole.

Dal punto vista interno, vengono in rilievo l'art. 110 del Codice (recentemente novellato dalla legge 29 aprile 2024, n. 56) e, in caso di futura approvazione, l'art. 8 (*Ricerca e sperimentazione scientifica nella realizzazione di sistemi di intelligenza artificiale in ambito sanitario*) del disegno di legge del Governo del 23 aprile 2024 contenente *Norme di principio in materia di intelligenza artificiale.*

L'art. 110, par. 1 indica due casi in cui **il consenso** dell'interessato **non è necessario** per il trattamento di dati sanitari.

1) La prima ipotesi prevista dall'articolo non costituisce, in realtà, una vera base giuridica di diritto interno *ex* art. 9, par. 2, lett. j, GDPR, limitandosi a richiamare quanto già esplicitato della norma europea. Pertanto, ad oggi, gli unici due indici normativi interni capaci di legittimare il trattamento in assenza di consenso sono: a) l'art. 12-*bis* del decreto legislativo 30 dicembre 1992, n. 502 (in base al quale il progetto di ricerca deve rientrare nel Piano sanitario nazionale stabilito dal Ministero della salute); b) l'art. 110-*bis*, co. 4, del Codice, il quale prevede un trattamento di favore per gli Istituti di ricovero e cura a carattere scientifico (IRCCS) per il trattamento ulteriore di dati personali per scopi di ricerca scientifica, in ragione della strumentalità dell'attività di assistenza sanitaria rispetto alla ricerca.

In riferimento a questa prima ipotesi, l'art. 110, co. 1, prima parte del Codice prescrive al titolare del trattamento l'adozione e la pubblicazione di una valutazione di impatto (VIP) ai sensi degli artt. 35 e 36 del GDPR.

2) La seconda ipotesi riguarda i casi in cui, a causa di particolari ragioni, contattare gli interessati per ottenere il consenso al trattamento risulta impossibile o implica uno sforzo sproporzionato, ovvero rischia di rendere

impossibile o di pregiudicare gravemente il conseguimento delle finalità della ricerca. Dunque, sul titolare del trattamento grava l'onere di provare le ragioni particolari ed eccezionali che, nel caso specifico, rendono il consenso "non necessario". In particolare, tali ragioni devono essere documentate nel progetto di ricerca, ai sensi del secondo paragrafo del punto 5.3 delle *Prescrizioni relative al trattamento dei dati personali effettuato per scopi di ricerca scientifica (aut. gen. n. 9/2016)*, GPDP, Provvedimento recante le prescrizioni relative al trattamento di categorie particolari di dati, ai sensi dell'art. 21, comma 1 del d.lgs. 10 agosto 2018, n. 101.

Tuttavia, in merito a questa seconda casistica di trattamento dei dati relativi alla salute in assenza di consenso dell'interessato, l'art. 110, primo comma, seconda parte del Codice impone al titolare del trattamento una serie di adempimenti sia sostanziali che procedurali: 1) l'adozione di misure appropriate per tutelare i diritti, le libertà e i legittimi interessi dell'interessato; 2) la sottoposizione del progetto di ricerca al competente comitato etico a livello territoriale per ottenerne un motivato parere favorevole; 3) l'osservanza di alcune garanzie che il Garante è chiamato a individuare ai sensi dell'art. 106, co. 2, lett. d), del Codice. In merito a tali garanzie, nel provvedimento del 9 maggio 2024 il Garante ha specificato che il titolare del trattamento è tenuto a motivare e documentare nel progetto di ricerca le ragioni etiche e/o organizzative a causa delle quali: a) risulta impossibile o sproporzionato informare gli interessati (e, quindi, acquisirne il consenso); oppure b) informare gli interessati rischia di rendere impossibile o di pregiudicare gravemente il conseguimento delle finalità della ricerca. Inoltre, al titolare del trattamento è richiesto di svolgere e pubblicare la valutazione di impatto (art. 35, GDPR), dandone comunicazione al Garante. In tale contesto, il Garante ha altresì promosso l'adozione di nuove Regole deontologiche per i trattamenti a fini statistici o di ricerca scientifica ai sensi degli artt. 2-*quater* e 106 del Codice.

L'attuale formulazione dell'art. 110 è frutto della novella apportata dalla L. 29 aprile 2024, n. 56, con cui è stato convertito il decreto-legge n. 19/2024 (c.d. Decreto PNRR). La precedente versione della norma richiedeva la consultazione preventiva del Garante per poter trattare i dati relativi alla salute in assenza di consenso dell'interessato. Operativamente, tale obbligo di consultazione ha rappresentato un oneroso ostacolo per numerosi enti di ricerca, indipendentemente dalla loro natura giuridica pubblica o privata e dalla presenza o meno di uno scopo di lucro, con progetti bloccati per lunghi periodi di tempo e spesso modificati nei loro obiettivi iniziali (soprattutto se aventi ad oggetto la costruzione di una biobanca o banca dati, ovvero lo sviluppo e l'implementazione di algoritmi).

Concludendo sul punto, la riforma dell'art. 110 del Codice, eliminando il requisito della preventiva consultazione del Garante, ha cercato di rispondere alle esigenze manifestate nel settore della ricerca scientifica e di superare così i *vulnera* economici e di tempo sopra menzionati.

Da ultimo va segnalato che, al netto delle due eccezioni brevemente esposte, nell'ordinamento italiano la regola generale per il trattamento dei dati per scopi di ricerca scientifica rimane il consenso dell'interessato e, in particolare, un tipo di consenso c.d. a fasi progressive, elaborato dal Garante attraverso una rigida interpretazione del Considerando n. 33 del GDPR. Pertanto, oltre al consenso iniziale, dovrebbe essere richiesto un ulteriore consenso specifico all'interessato

non solo per eventuali e futuri progetti di ricerca, ma anche per ulteriori fasi del medesimo progetto, definibili al momento della raccolta del consenso soltanto in via generale.

Il panorama giuridico interno potrebbe mutare radicalmente in caso di futura approvazione del disegno di legge del Governo del 23 aprile 2024 (*Norme in materia di intelligenza artificiale*), il quale permetterebbe, nel settore in esame, di superare in via generale il requisito del consenso dell'interessato, andando a costituire una disposizione di diritto interno ai sensi dell'art. 9 par. 2, lett. g, GDPR. Infatti, il primo comma dell'art. 8 (*Ricerca e sperimentazione scientifica nella realizzazione di sistemi di intelligenza artificiale in ambito sanitario*) dichiara di rilevante interesse pubblico (*ex* art. 9 par. 2, lett. g, GDPR) i trattamenti eseguiti da soggetti pubblici e privati senza scopo di lucro per la ricerca e la sperimentazione scientifica nella realizzazione di sistemi di IA per finalità di «*prevenzione, diagnosi e cura di malattie, sviluppo di farmaci, terapie e tecnologie riabilitative, realizzazione di apparati medicali, incluse protesi e interfacce fra il corpo e strumenti di sostegno alle condizioni del paziente, di salute pubblica, incolumità della persona, salute e sicurezza sanitaria, in quanto necessari ai fini della realizzazione e dell'utilizzazione di banche dati e modelli di base*». In altre parole, la novella riconoscerebbe, per i fini menzionati, la possibilità del trattamento secondario dei dati personali (anche appartenenti alle categorie particolari e, dunque, anche i dati sanitari) in assenza del consenso dell'interessato, fermo restando l'obbligo di informativa (**«*che può essere assolto anche mediante messa a disposizione di un'informativa generale sul sito web del titolare del trattamento*»**, art. 8, co. 2) e l'adozione di una serie di condizioni preventive (art. 8, co. 3). In particolare, i trattamenti dovranno essere approvati dai comitati etici interessati e dovranno essere comunicati al Garante. Trascorsi trenta giorni da tale comunicazione, in assenza di un esplicito provvedimento di blocco da parte di quest'ultimo, il trattamento potrà essere legittimamente intrapreso.

Allo stato dell'arte, il disegno di legge citato è ancora in fase di esame parlamentare, tuttavia è opportuno dar conto della posizione critica espressa dal Garante. In particolare, il quest'ultimo ha precisato che, a suo dire, sarebbe necessario: 1) eliminare il riferimento alla possibilità di assolvere l'obbligo di informativa in forma generale, con pubblicazione sul sito web del titolare, non essendo questa modalità «*compatibile con tale ipotesi di uso secondario dei dati*»; 2) conformare il co. 1 ai requisiti di determinatezza di cui agli artt. 6, par. 3, lett. b), 9, par. 2, lett. g, GDPR e 2-*sexies* del Codice; 3) prevedere, con riferimento all'uso secondario dei dati, le garanzie di cui all'art. 89 GDPR (cfr. Parere su uno schema di disegno di legge recante disposizioni e deleghe in materia di intelligenza artificiale, 2 agosto 2024, n. 10043532).

3) *Un paziente può ritirare il consenso all'utilizzo dei suoi dati (ancorché, naturalmente, privati di ogni informazione che permetta il riconoscimento della sua identità personale) una volta che questi siano stati utilizzati p.es. per costruire una biobanca o generare un modello diagnostico/predittivo basato su IA?*

Anche la risposta a questa domanda dipende dal tipo di dato a cui si fa riferimento. Qualora i dati siano anonimi, il paziente non è titolare di alcun diritto di revoca, in quanto, come visto, tali dati non rientrano nell'alveo applicativo del

GDPR. Di converso, nel caso in cui i dati siano pseudonimizzati, è sempre possibile, ai sensi dell'art 7, par 3 del GDPR, esercitare il diritto di revoca, a patto che il trattamento fosse basato sul consenso dell'interessato (in questa circostanza del paziente). Il medesimo paragrafo specifica, però, che la revoca ha efficacia esclusivamente ***pro futuro,*** senza pregiudicare la liceità del trattamento già realizzato in base al consenso.

In caso di revoca del consenso da parte dell'interessato, il titolare deve immediatamente interrompere le attività di trattamento e, in assenza di altra base giuridica che giustifichi la conservazione dei dati per ulteriori trattamenti, i dati personali devono essere cancellati (*cfr.* GPDP, Provvedimento 12 ottobre 2023, n. 9953841). Al contrario, il trattamento dei dati personali potrebbe proseguire laddove il titolare del trattamento individuasse un'altra base giuridica idonea (*ex* artt. 6 e 9, GDPR).

Nell'evenienza di base giuridica diversa dal consenso, il GDPR riconosce all'interessato il diritto di opporsi al trattamento. A differenza della revoca, la quale non necessità di alcuna giustificazione, il diritto di opposizione richiede un'adeguata motivazione, legata alla "situazione particolare" dell'interessato (art. 21, GDPR). Anche per questo, il titolare può superare l'opposizione, dimostrando «*l'esistenza di motivi legittimi cogenti per procedere al trattamento che prevalgono sugli interessi, sui diritti e sulle libertà dell'interessato*» (art. 21, GDPR). In riferimento alla ricerca scientifica, il sesto comma dell'art. 21, GDPR sancisce, inoltre, che il diritto di opposizione ha come limite l'interesse pubblico eventualmente sotteso al trattamento.

All'atto pratico, spetta all'informatica chiarire quale sia l'impatto che la revoca o l'opposizione hanno sul funzionamento dell'algoritmo: una volta che il dato è stato utilizzato nel processo di apprendimento, la macchina può disimparare?

4) *Per qualsiasi procedura medica diagnostica o terapeutica che possa comportare un rischio per il paziente (p.es. indagine di tomografia computerizzata, che comporta esposizione a radiazioni ionizzanti; procedure interventistiche guidate o assistite da IA), è richiesta la compilazione di un consenso informato. Il paziente dovrà essere adeguatamente informato in tutti i casi in cui algoritmi di IA siano utilizzati (p.es. miglioramento della qualità dell'immagine), o soltanto in quei casi in cui l'output dell'algoritmo possono avere effetti diretti e sostanziali sul management del paziente stesso (p.es. strumenti di supporto decisionale)?*

Ai sensi dell'art. 1, co. 4 della legge n. 219/2017, per procedere a qualsiasi trattamento sanitario, più o meno rischioso, è necessario ottenere uno specifico consenso informato del paziente, da documentare, salvo rari casi, in forma scritta.

Pertanto, a ben vedere, la domanda posta attiene al contenuto del consenso, ossia alle informazioni concretamente fornite al paziente al momento del consenso informato. Il paziente deve essere informato ogni volta che l'IA viene utilizzata in una qualsiasi fase del trattamento sanitario, indipendentemente dallo scopo? Oppure è necessario informarlo solo se l'uso dell'IA comporta un rischio per la sua salute?

Le fonti di riferimento – oltre all'art.32, co. 2, Cost. – sono l'art. 1, co. 3 della già citata legge n. 219/2017 e l'art. 33 del Codice di deontologia medica. Dal combinato disposto delle due norme da ultimo citate emerge che il paziente debba

essere informato in modo completo, aggiornato e comprensibile, tra le altre cose, dei rischi del trattamento sanitario, mentre non è espressamente richiesto che il contenuto dell'informativa si estenda anche a tutte le tecniche e agli strumenti utilizzati nel trattamento.

Quindi, il paziente, compatibilmente con le sue concrete possibilità cognitive e con le conoscenze del medico, dovrà, in genere, essere informato qualora l'impiego di strumenti di IA comporti, oltre a benefici, anche rischi per la sua salute. In particolare, egli dovrà essere reso edotto degli eventuali diversi e nuovi rischi introdotti della tecnologia. Inoltre, può ritenersi, quantomeno per fini precauzionali, che egli debba essere informato anche quanto l'IA svolge un ruolo centrale nel trattamento sanitario e nella scelta terapeutica. Di converso, almeno a un'interpretazione letterale, non sembra ci sia alcun obbligo del medico di informare il paziente della circostanza che l'intelligenza artificiale intervenga nel trattamento senza incidere sostanzialmente sullo stesso e senza comportare rischi apprezzabili per la salute. D'altronde, allo stesso modo, oggi il paziente non viene di certo reso edotto dell'utilizzo e del funzionamento di ogni dispositivo tecnico di cui si avvale il medico nel corso del trattamento.

A supporto di questa ricostruzione, la dottrina, facendo anche leva su studi di psicologia cognitiva e citando un importante orientamento della Corte di cassazione (Cass. civ., 20 maggio 2016, n. 10414; Cass. civ., 19 settembre 2014, n. 19731; Cass. civ., 11 dicembre 2013, n. 27751; Cass. civ., 30 luglio 2004, n. 14638), rileva l'irragionevolezza di informare i pazienti di ogni dettaglio tecnico relativo al trattamento. Più precisamente, spesso la conoscenza degli elementi tecnici del trattamento sanitario non ha alcuna reale incidenza positiva sulla scelta del paziente e sulla tutela del bene salute, se non, potenzialmente, in chiave negativa. Infatti, si evidenzia che sovraccaricare il paziente di informazioni tecniche può confonderlo e, soprattutto, alimentare scelte irrazionali, basate sulla diffidenza verso la tecnologia o sulla sopravvalutazione del verificarsi di c.d. rischi "anomali". Cionondimeno, all'opposto, viene evidenziato che, qualora sia il paziente a porre specifiche domande sulle tecnologie utilizzate, il medico dovrà fornirgli (senza che questo comporti sforzi irragionevoli) adeguate e complete risposte in merito all'impiego dell'IA. Infatti, per quanto l'eventuale scelta di celare un'informazione richiesta possa essere motivata dalla volontà di proteggere il paziente da decisioni irrazionali, tale pratica rifletterebbe una concezione ormai superata e paternalistica della medicina, finendo per compromettere il rapporto di fiducia tra medico e paziente.

Inoltre, nonostante un funzionamento del tutto autonomo della macchina sia lontano dalla realtà e forse nemmeno auspicabile, qualora la decisione terapeutica fosse interamente presa dal dispositivo medico intelligente troverebbe applicazione l'art. 22 del GDPR, ai sensi del quale *l'interessato ha il diritto di non essere sottoposto a una decisione basata unicamente sul trattamento automatizzato [...] che produca effetti giuridici che lo riguardano o che incida in modo analogo significativamente sulla sua persona*. Pertanto, in questo scenario, il consenso del paziente rappresenterebbe un passaggio inevitabile.

Anche in tale frangente il quadro appena esposto potrebbe cambiare radicalmente nel prossimo futuro. Infatti, l'art. 7, co. 3 del menzionato disegno di legge sull'IA garantisce inequivocabilmente all'interessato il diritto di essere informato in relazione a tre aspetti: 1) l'utilizzo di tecnologie di IA in ambito

sanitario; 2) i vantaggi, in termini diagnostici e terapeutici, derivanti dall'utilizzo di tali tecnologie; 3) le logiche decisionali utilizzate. Alla luce di questo, risulta evidente che, in caso di definitiva approvazione del DDL, l'obbligo informativo si estenderebbe fino a ricomprendere in via generale l'impiego di strumenti di IA, indipendentemente dagli eventuali rischi per la salute del paziente e anche dall'essere o meno il dispositivo di IA utilizzato come supporto per la decisione del medico.

Da ultimo, può essere interessante evidenziare come il citato articolo del disegno di legge italiano sia in linea con il dettato dell'art. 4001-3 del *Code de la santé publique* francese.

DA PARTE DEGLI INFORMATICI:

5) *Come si collocano, all'interno della regolamentazione attuale, strumenti di IA decisionali interattivi, capaci di riconoscere i propri limiti demandando all'uomo la decisione su dei casi in cui il modello si renda conto di non poter essere abbastanza accurato? Potrebbero avere un impatto rispetto alla identificazione delle responsabilità?*

Al momento, simili strumenti di IA non sono contemplati nell'ordinamento; tuttavia, una chiara indicazione normativa favorevole allo sviluppo di sistemi "consci dei propri limiti" la si potrebbe rinvenire nell'art. 13, comma 1, dell'AI Act, ai sensi del quale: «*I sistemi di IA ad alto rischio sono progettati e sviluppati in modo tale da garantire che il loro funzionamento sia sufficientemente trasparente da consentire agli utenti di interpretare l'output del sistema e utilizzarlo adeguatamente*».

Peraltro, da un punto di vista pratico, l'autovalutazione di accuratezza da parte delle IA potrebbe avere un impatto decisivo in uno dei momenti più critici del rapporto medico-IA, cioè quello dell'emersione del dissenso (vedi la risposta *sub* 1): è evidente che tanto minore dovesse risultare l'accuratezza (autodichiarata) della IA, tanto più potrebbe apparire ragionevole, e scusabile, la confidenza del medico nel proporre una diagnosi difforme, senza il timore di subire automatismi giudiziari in caso di errore (e viceversa).

Bibliografia

N. Amore, E. Rossero, *Robotica e intelligenza artificiale nell'attività medica. Organizzazione, autonomia, responsabilità*, Il Mulino, 2023.

A. Bertolini, *Dall'imaging ai sistemi esperti: la responsabilità del medico e le nuove frontiere della medicina difensiva*. M. Faccioli (a cura di), Profili giuridici dell'utilizzo della robotica e dell'intelligenza artificiale in medicina, ESI, 2022, 39 ss.

A. Bertolini., I*nsurance and Risk Management for Robotic Devices: Identifying the Problems*. Global Jurist, 2016, 291 ss.

A. Bertolini, *Robotic prostheses as products enhancing the rights of people with disabilities. Reconsidering the structure of liability rules*. International Review of Law, Computers & Technology, 2015, 116 ss.

A. Bertolini, *Artificial intelligence does not exist! defying the technology- neutrality narrative in the regulation of civil liability for advanced technologies*. Europa e diritto privato, 2022, 369 ss.

V. Calderai, *Consenso informato*. Enciclopedia del diritto. Annali VIII, 2015, 225 ss.

J. Chen, E.S. Dove, H. Bhakuni, *Explicit consent and alternative data protection processing grounds for health research*. E. Kosta, R. Leenes (a cura di), Research Handbook on EU Data Protection Law, Elgar, 2022, 474 ss.

A. Cioni, *Nuovi pregi e vecchi difetti della proposta di direttiva sulla responsabilità da prodotto difettoso, con particolare riferimento all'onere della prova*. Responsabilità civile e previdenza, 2, 2023, 656 ss.

A. Colaruotolo, *Intelligenza artificiale e responsabilità medica: novità, continuità e criticità*. Responsabilità medica, 3, 2022, 299 ss.

E. Colletti, *Intelligenza artificiale e attività sanitaria. Profili giuridici dell'utilizzo della robotica in medicina*. Rivista di diritto dell'economia, dei trasporti e dell'ambiente, 2021, 201 ss.

G. Comandè, G. Schneider, *Differential Data Protection Regimes in Data-Driven Research: Why the GDPR is More Research-Friendly Than You Think*. German Law Journal, 2022, 559 ss.

M.T. Contaldo, G. Pasceri, G. Vignati, L. Bracchi, S. Triggiani, G. Carrafiello, *AI in Radiology: Navigating Medical Responsibility*. Diagnostics, 2024, 14, 1506 ss.

C. De Menech, *Intelligenza artificiale e autodeterminazione in materia sanitaria*. M. Faccioli (a cura di), Profili giuridici dell'utilizzo della robotica e dell'intelligenza artificiale in medicina, ESI, 2022, 9 ss.

V. Di Gregorio, *Intelligenza artificiale e responsabilità civile: quale paradigma per le nuove tecnologie?* Danno e responsabilità, 1, 2022, 61 ss.

European Society of Radiology (ESR). *The new EU General Data Protection Regulation: what the radiologist should know*. Insights Imaging, 8, 2017, 295 ss.

M. Faccioli, *Intelligenza artificiale e responsabilità sanitaria*. La nuova giurisprudenza civile commentata, 3, 2023, 732 ss.

A. Fiorentini, Machine learning *e dispositivi medici: riflessioni in materia di responsabilità civile*. Il corriere giuridico, 10, 2021, 1258 ss.

L. Georgieva, C. Kuner, *Article 9. Processing of special categories of personal data* in C. Kuner. L.A. Bygrave e C. Docksey (a cura di), The EU General Data Protection Regulation (GDPR), OUP, 2020, 365 ss.

A.G. Grasso, *Diagnosi algoritmica errata e responsabilità medica*. Rivista di diritto civile, 2, 2023, 334 ss.

P. Guarda, *Ricerca medica, biomedica ed epidemiologica*. R. D'Orazio, G. Finocchiario, O. Pollicino, G. Resta (a cura di), Codice della privacy e data protection, Giuffrè, 2021, 1369 ss.

C. Iagnemma, *I 'robot medici': profili problematici in tema di alleanza terapeutica e di responsabilità penale*. Corti supreme e salute, 2, 2020, 441 ss.

F.C. La Vattiata, *Artificial intelligence in Healthcare: Risk Assessment and Criminal Law*. Diritto penale e uomo, 12, 2020, 23 ss.

G. Maliha, S. Gerke, I.G. Cohen, R.B. Parikh, *Artificial Intelligence and Liability in Medicine: Balancing Safety and Innovation.* Milbank Q., 2021, 99(3), 629 ss.

E. Palmerini, *AI Systems and the Issue of Liability in the European and National Regulatory Strategies.* P. Morgan (a cura di), Tort Liability and Autonomous Systems Accidents. Common and Civil Law Perspectives, Edward Elgar, 2023, 63 ss.

M. Pruski, *AI-Enhanced Healthcare: Not a new Paradigm for Informed Consent.* Bioethical Inquiry, 2024, 1 ss.

S. Quattrocolo, *Intelligenza artificiale e giustizia: nella cornice della Carta Etica Europea, gli spunti per un'urgente discussione tra scienze penali e informatiche.* La legislazione penale, 18.12.2018, 1 ss.

U. Ruffolo, *L'intelligenza artificiale in sanità: dispositivi medici, responsabilità e "potenziamento".* Giurisprudenza italiana, 2, 2021, 456 ss.

D. Schönberger, *Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications.* International Journal of Law and Information Technology, 2019, 171 ss.

P. Verbruggen and B. Van Leeuwen, *The liability of notified bodies under the EU's new approach: the implications of the PIP breast implants case.* European law review, 2018, 394 ss.

G. Votano, *Intelligenza artificiale in ambito sanitario: il problema della responsabilità civile.* Danno e responsabilità, 6, 2022, 669 ss.

G. Wagner, *Next Generation EU Product Liability - For Digital and Other Products.* Journal of European Tort Law, 2024, 172 ss.

AMBITO MEDICO

Profili applicativi

NOZIONI INTRODUTTIVE:

Nel 1959 Arthur Samuel introdusse il concetto di "apprendimento automatico" (*machine learning*, ML), inteso come una classe di sistemi di IA debole che prendono decisioni utilizzando modelli costruiti a partire dai dati inseriti per l'addestramento. In altri termini, un sistema basato su ML può fare previsioni su nuovi dati a seguito di un precedente addestramento, senza la necessità di essere appositamente programmato o di ricorrere a modelli predefiniti. Una conseguenza notevole di tale proprietà è che le prestazioni di un sistema di ML tendono ad aumentare all'aumentare dell'esperienza del sistema stesso.

Nel ML classico i dati vengono etichettati da esperti umani e categorizzati in base alle loro proprietà utilizzando metodi statistici. Affinché un algoritmo di ML possa riprodurre con successo l'analisi di un'immagine (p.es. una radiografia del torace) da parte di un Radiologo umano, dovrà essere addestrato a partire non da un unico *dataset* di apprendimento non etichettato (contenente molte tipologie eterogenee di anomalie radiografiche), ma da diversi *dataset* che rafforzano l'apprendimento per ciascuna classe di reperti anormali (p.es. cardiache, mediastiniche, polmonari, ossee) e da *dataset* aggiuntivi per varie sottoclassi di anomalie (p.es. cardiopatie congenite).

Il ML costituisce la massima espressione della potenza di calcolo di un sistema informatico, intesa come la sua capacità di elaborare indefinitamente dati senza affaticarsi. Tuttavia, in un sistema potente ma mal controllato potrebbero verificarsi problematiche come il sovra-adattamento (*overfitting*), che comporta un peggioramento delle prestazioni dell'algoritmo su nuovi dati a seguito dell'inserimento nel modello di apprendimento (unitamente ai dati di ingresso corretti e in maniera indistinguibile da essi) di fluttuazioni statistiche non significative dei dati immessi. Nell'ambito dell'analisi delle immagini, il fenomeno del sovra-adattamento può essere amplificato dalla possibilità di varianti anatomiche non patologiche (come ossa accessorie, oppure strutture congenitamente assenti o ipoplasiche).

All'estremo opposto si colloca il fenomeno del sotto-adattamento (*underfitting*), che sussiste quando un algoritmo di AA non è in grado di riprodurre fedelmente (p.es. a causa di un insufficiente addestramento) le relazioni esistenti tra le caratteristiche di un *dataset* e una data variabile, interpretando così i dati in ingresso in maniera eccessivamente semplificata.

L'apprendimento profondo (in inglese *deep learning*, DL) rappresenta un sottoinsieme del ML, che può consentire di ottenere prestazioni migliori rispetto al ML classico. Anziché ricorrere a forme di etichettatura o di ingegnerizzazione delle caratteristiche, gli algoritmi di DL apprendono in maniera autonoma le caratteristiche più adatte per classificare i dati forniti in ingresso, in base allo specifico compito assegnato, utilizzando reti neurali artificiali di ispirazione biologica. Le reti neurali convoluzionali (in inglese: *Convolutional Neural Networks*,

CNN) sono un tipo particolare di rete neurale ottimizzata per il riconoscimento di *patterns* all'interno delle immagini e rappresentano l'approccio più comune per l'elaborazione delle immagini biomediche, potendo accettare in ingresso immagini bi- o tridimensionali.

Il DL si è dimostrato promettente per l'estrazione di caratteristiche dalle immagini biomediche. In queste applicazioni, le unità computazionali sono definite come livelli strutturati in maniera gerarchica e fra loro integrati per estrarre le caratteristiche intrinseche delle immagini: attraverso una rete neurale convoluzionale, un sistema di DL può p.es. estrarre le caratteristiche intrinseche di una neoplasia per fornire informazioni prognostiche, con un evidente impatto sulla gestione clinica del paziente.

La potenza di calcolo delle infrastrutture hardware e la disponibilità di dati adeguati all'apprendimento sono due fattori che condizionano direttamente la funzionalità delle reti neurali impiegate per il DL. Se la potenza di calcolo aumenta significativamente nell'arco di anni o mesi e può essere pertanto limitata a breve termine solo da fattori temporali o economici, la soluzione al problema di una scarsa disponibilità di *dataset* strutturati per l'addestramento non è banale e può ostacolare lo sviluppo e la diffusione su vasta scala di tali sistemi.

PROBLEMATICHE APPLICATIVE:

L'interpretazione automatica delle immagini rappresenta una delle applicazioni radiologiche dell'IA che desta maggior entusiasmo. È ormai disponibile un'ampia letteratura scientifica che attesta l'efficacia dell'IA nel supportare Radiologi umani in svariate procedure diagnostiche, come (per citarne solo alcune) la diagnosi di polmonite e di tubercolosi, l'individuazione di fratture e la stima dell'età ossea. Altre applicazioni di IA consentono di ottimizzare i protocolli di acquisizione delle immagini (consentendo di ottenere immagini diagnostiche in tempi minori e con una minor dose di radiazioni ionizzanti e mezzo di contrasto) o indicare la priorità della refertazione di taluni esami (p.es. TC cranio in urgenza in pazienti con ictus emorragico), migliorando i flussi di lavoro ed evitando ritardi diagnostici, soprattutto in situazioni di maggior impegno del personale radiologico. Tuttavia, attualmente non sono disponibili soluzioni IA validate e disponibili per uso commerciale in grado di interpretare le immagini e generare un referto in maniera autonoma.

La disponibilità di dati strutturati e categorizzati su scala multicentrica (*big data*), ottenibili attraverso l'integrazione con i sistemi RIS/PACS, è un requisito essenziale per lo sviluppo e la diffusione dell'IA in ambito radiologico. Tuttavia, a tutt'oggi questa esigenza (vitale per un addestramento efficace degli algoritmi di IA) si scontra spesso con il fatto che la maggior parte dei referti radiologici è scritta sotto forma di testo narrativo non strutturato, pertanto non direttamente fruibile per l'estrazione automatizzata delle informazioni in essi contenute. Questo problema potrebbe essere risolto con l'adozione del referto strutturato (RS), che, se adeguatamente implementato e supportato dall'utilizzo di algoritmi di elaborazione del linguaggio naturale (*Natural Language Processing* o NLP, anch'essi basati su IA), può favorire la raccolta e l'elaborazione di *big data* attraverso una semantica e un lessico standardizzati, riducendo inoltre il rischio di

errori e rendendo possibili ricerche automatizzate per la costruzione di *trials* multicentrici.

Un campo di ricerca di grande attualità connesso all'IA è costituito dalla radiomica, definita come l'estrazione di una grande mole di caratteristiche quantitative dalle immagini biomediche convenzionali, al fine di ottenere dati utilizzabili in sistemi di supporto clinico-decisionale per migliorare l'accuratezza diagnostica, prognostica e predittiva. La radiomica si avvale di sistemi di IA per cercare correlazioni tra biomarcatori di imaging e altri biosegnali (come dati clinici, parametri di laboratorio, etc.), fornendo informazioni non ottenibili con la semeiotica radiologica standard, come quelle relative alla risposta precoce ai trattamenti, la predizione dell'aggressività biologica di una neoplasia, l'esistenza di bersagli molecolari per eventuali terapie mirate, la previsione della prognosi individuale e la personalizzazione delle terapie. La radiomica può comprendere l'analisi dell'assetto genetico di un tessuto, nel qual caso si parla di radiogenomica. Le caratteristiche ("firme", *signatures*) radiomiche possono essere calcolate utilizzando programmi dedicati a partire da immagini acquisite con protocolli di routine. Tuttavia, per quanto la radiomica rivesta grande interesse in virtù delle proprie potenzialità, al momento il suo utilizzo è tuttora limitato alla pura ricerca da diversi fattori, che comprendono la frequente mancanza di validazione su base multicentrica, la scarsa "spiegabilità" (*explainability*) e universalità dei metodi utilizzati e la difficoltà di garantire un'adeguata qualità e omogeneità dei dati inseriti per ottenere risultati riproducibili, anche a causa delle differenze tra i protocolli di acquisizione di immagini acquisite con apparecchiature diverse o in centri diversi.

Quesiti

DA PARTE DEI GIURISTI:

1) *Che differenza si riscontra, dal punto di vista sia metodologico che operativo, quando si deve realizzare un atto medico in un settore "coperto da linee guida" (o buone pratiche), rispetto a uno direttamente suggerito per il determinato caso clinico da un dispositivo medico IA?*

In entrambi gli scenari posti dal quesito, la responsabilità del medico è quella di fornire al paziente il percorso diagnostico-terapeutico più appropriato e conforme agli standard professionali. La principale differenza, in tal caso, sta nel processo decisionale alla base di tale percorso. Le linee guida forniscono un insieme di raccomandazioni basate su evidenze scientifiche per la gestione di determinati scenari clinici. Da un punto di vista prettamente metodologico, il medico segue pertanto un approccio basato sull'evidenza (i.e., *evidence based medicine*), integrando le linee guida nella valutazione del paziente. Operativamente, il medico esercita la propria competenza professionale per adattare le raccomandazioni, generiche, alle circostanze del caso clinico, specifiche. Diverso è il caso del "suggerimento" fornito dal dispositivo medico IA, che è specificatamente formulato per il singolo caso. Al medico è comunque richiesto di interpretare tale suggerimento, integrandolo criticamente nel processo decisionale, tenendo in considerazione anche altre informazioni rilevanti, alla luce delle condizioni specifiche del paziente, della

propria esperienza clinica, e, ovviamente, anche delle linee guida. Pertanto, la competenza professionale del medico rimane centrale nel processo decisionale e nell'erogazione delle cure.

2) *Quali sono i possibili approcci sanitari ipotizzabili, anche alla luce della metodologia diagnostica consolidata, per gestire in modo standardizzato e affidabile le procedure per effettuare una diagnosi per immagini in contesti sanitari dotati di sistemi di IA?*

I sistemi di IA "di ausilio intellettuale" supportano il medico nel processo decisionale e nell'atto medico, che, nella diagnostica per immagini, è rappresentato, in buona parte, dall'interpretazione di un'immagine radiologica al fine di raggiungere una diagnosi. In tal senso, i sistemi di IA potrebbero configurarsi come un "*second reader*", in grado di fornire una propria, per quanto "asettica", opinione. Nel più semplice modello di integrazione, il medico radiologo decide se e quando interpellare l'output del sistema di IA direttamente sul PACS, per il raggiungimento della diagnosi.

Tale tipo di approccio tuttavia risulterebbe poco standardizzato, variando di caso in caso e di radiologo in radiologo. Diversi approcci sono stati proposti dalla comunità scientifica per vari scenari clinici, per lo più attraverso studi retrospettivi. Uno degli ambiti nel quale sono stati proposti diversi approcci per l'integrazione di sistemi di IA nel *workflow* è la radiologica senologica. Dembrower et al. hanno proposto un esempio di integrazione di un sistema di IA, con la duplice funzione di *rule-out* e *rule-in* sulla base del punteggio assegnato dal sistema stesso.

Stanno inoltre emergendo i primi studi prospettici, tra cui quello di Ng et al., sempre nello stesso ambito, in cui l'IA veniva integrata insieme a un *reader* addizionale (sempre nell'ambito dello screening mammografico) per rivalutare tutti i casi negativi. Tuttavia, allo stato attuale, risulta difficile individuare dei modelli standard relativamente alle strategie di integrazione. Esistono infatti diverse fonti di variabilità, indipendentemente dallo scenario clinico applicativo. In ogni caso, l'integrazione di un sistema di IA non può chiaramente prescindere dalla formazione del personale coinvolto e dalla realizzazione di sistemi di monitoraggio continuo della performance, per garantire l'accuratezza e l'affidabilità nel tempo del sistema di IA, con la realizzazione di meccanismi di feedback che coinvolgano il personale nell'ottimizzazione delle prestazioni dei sistemi di IA.

3) *Quali sono i possibili approcci sanitari per gestire in modo standardizzato e affidabile l'impiego tecnico-operativo dei sistemi di IA nella diagnostica per immagini?*

Attualmente, non esistono standard relativi all'impiego operativo dei sistemi di IA nella diagnostica per immagini. Tuttavia, prettamente da un punto di vista tecnico, è necessario che l'IA si integri e sia in grado di dialogare con gli standard attualmente vigenti e ampiamente consolidati, come DICOM (*Digital Imaging and Communications in Medicine*) e HL7 (*Health Level Seven*). DICOM è uno standard internazionale per la gestione, lo scambio e la visualizzazione di informazioni mediche, specialmente immagini radiologiche come radiografie, immagini ecografiche, immagini di tomografie computerizzate e di risonanza magnetica. Questo standard fornisce una struttura comune per l'archiviazione e la

trasmissione di dati, garantendo l'interoperabilità tra dispositivi e sistemi di imaging di diversi produttori. HL7, d'altra parte, è uno standard che regola lo scambio di dati clinici e amministrativi tra sistemi informativi sanitari. Questo standard facilita la comunicazione tra diverse applicazioni mediche, consentendo lo scambio di informazioni cruciali per la diagnosi e il trattamento dei pazienti. Integrare l'IA nella diagnostica per immagini in conformità con questi standard non solo garantisce l'interoperabilità e la coerenza dei dati, ma facilita anche lo scambio di informazioni tra professionisti sanitari e istituzioni, migliorando complessivamente l'efficienza e la qualità delle cure mediche. Inoltre, l'adesione a tali standard favorisce lo sviluppo di soluzioni IA che possono essere implementate in modo più fluido e scalabile all'interno degli ambienti sanitari esistenti.

4) *Quali sono i possibili approcci sanitari per gestire in modo standardizzato e affidabile le modalità di espressione del dissenso da parte del medico radiologo rispetto alle diagnosi formulate dalle IA?*

Non esistono attualmente standard in merito alla formulazione del dissenso da parte del medico radiologo rispetto alla diagnosi formulata dalle IA. Verrebbe anche da chiedersi se tale dissenso debba essere espresso, o addirittura giustificato, in ogni occasione in cui tale dissenso venisse riscontrato. Di fatti, nell'atto della refertazione, il radiologo raggiunge una conclusione diagnostica alla luce di determinate evidenze, o, in taluni casi, nonostante alcune evidenze, cliniche o radiologiche che siano (e.g. un nodulo polmonare con caratteristiche morfologiche benigne, in un soggetto non-fumatore ed a basso rischio, potrebbe essere considerato maligno in caso di accrescimento dimensionale). Tuttavia, un conto è formulare una conclusione alla luce di evidenze, un conto è la formulazione di tale conclusione alla luce o nonostante una inferenza formulata da un algoritmo (tra l'altro, non trasparente) derivante, almeno in teoria, dalle stesse evidenze (di fatti, gli algoritmi sono "addestrati" su dati clinici o immagini, le stesse informazioni valutate dal radiologo nel corso del suo atto diagnostico). Per quanto tale dissenso possa rilevarsi rilevante in caso di errore diagnostico, ci si può aspettare anche un certo numero di casi in cui sia il sistema di IA a sbagliarsi. La registrazione di ogni dissenso potrebbe comportare una riduzione della *confidence* diagnostica e un aumentato timore di eventuali ripercussioni legali, con incremento delle pratiche di medicina difensiva.

5) *Nel formulare una diagnosi, quanto il timore di poter essere chiamati in giudizio dal paziente potrebbe incidere sulla scelta di discostarsi dal verdetto espresso dall'IA?*

Nella pratica clinica radiologica, dato un quesito, non è infrequente che le evidenze cliniche e radiologiche siano in disaccordo. In tal caso, con spirito critico, il medico radiologo deve soppesare le diverse evidenze per raggiungere la diagnosi corretta. L'IA aggiunge un ulteriore elemento a questo quadro, in quanto l'output fornito dal sistema può porsi in accordo o in disaccordo con la conclusione diagnostica raggiunta dal radiologo sulla base delle evidenze sovradescritte. Il medico radiologo dovrà pertanto soppesare anche questo elemento nel processo decisionale, correndo il rischio di potere essere chiamato in giudizio qualora la scelta di discostarsi dal verdetto espresso dall'IA risultasse errata. La medicina difensiva è già, purtroppo, una realtà, con conseguenze economiche e sociali

rilevanti e tutt'ora di difficile risoluzione. L'utilizzo di un approccio difensivo da parte del medico che utilizza l'IA, potrebbe portare ad un ulteriore incremento del carico di lavoro e di esami di II livello richiesti per approfondimenti diagnostici, creando ulteriori problemi là dove l'IA era stata invece implementata per risolverli. Tale timore potrebbe avere conseguenze nefaste là dove l'IA è applicata con poco spirito critico e senza una adeguata conoscenza della materia radiologica, in particolare in chi ancora non abbia sviluppato abbastanza esperienza. Una adeguata formazione del personale è essenziale per limitare queste problematiche.

DA PARTE DEGLI INFORMATICI:

6) *Strumenti decisionali interattivi, capaci di riconoscere i propri limiti demandando all'uomo la decisione su dei casi in cui il modello si renda conto di non poter essere abbastanza accurato, potrebbero essere accettati? O porterebbero a maggiore sfiducia?*

La trasparenza appare sempre di più come un elemento imprescindibile e cruciale, fortemente stressato anche dall'ormai noto AI act. Affinché il radiologo possa soppesare nel modo più critico possibile l'output del sistema, è necessario che quest'ultimo possa essere "interpretato". Là dove non sia possibile comprendere l'effettivo processo che ha portato il sistema a fornire quell'output, l'espressione quantomeno di un "grado di confidenza" da parte del modello potrebbe essere di aiuto. Per questo motivo, modelli che demandino al radiologo casi in cui l'accuratezza non è ottimale, sarebbero meritevoli di maggior fiducia rispetto a modelli che forniscano esclusivamente output binari.

7) *Se per il design di un sistema di IA si seguisse un modello di co-designing, potrebbe questo fornire più garanzie e portare i medici a una maggiore fiducia nei sistemi?*

La risposta è assolutamente positiva. Il coinvolgimento di più figure professionali nel design di un modello diagnostico è di fondamentale importanza per la realizzazione di modelli non solo efficienti, ma anche utili ed efficaci.

8) *Avere a disposizione degli strumenti per una valutazione del rischio di privacy di tipo data-driven potrebbe migliorare la diffusione dei sistemi di AI?*

Certamente, i sistemi di valutazione del rischio di privacy devono andare di pari passo con l'impiego dei sistemi di IA. Nell'era dei big data, i dati sono ormai dei beni che vanno tutelati, in particolar modo in ambito sanitario. Spingendoci oltre rispetto al quesito, solo se effettivamente tale prerequisito venisse rispettato, ci si potrebbe aspettare una diffusione dei sistemi di IA, e in particolare un loro utilizzo in ambito sanitario: senza tale prerequisito, ciò non sarebbe affatto possibile.

Bibliografia

J. Bajva, *Artificial intelligence in healthcare: transforming the practice of medicine.* Future Healthcare Journal, 2021.

A.P. Brady, E. Neri, *Artificial Intelligence in Radiology, Ethical Considerations.* Diagnostics, 2020.

F. Coppola, L. Faggioni *et al.*, *Human, All Too Human? An All-Around Appraisal of the "Artificial Intelligence Revolution" in Medical Imaging*. Frontiers in Psychology, 2021.

European Society of Radiology. *What the radiologist should know about artificial intelligence - an ESR white paper*. European Radiology. Insights Into Imaging, 2019.

K. Dembrower *et al.*, *Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study*. Digital Health, 2020, 468 ss.

E. Dikici *et al.*, *Integrating AI into radiology workflow: levels of research, production, and feedback maturity*. Journal of Medical Imaging, 2020, 1 ss.

J. R. Geis *et al.*, *Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement*. Radiology, 2019.

M. Klontzas, S.C. Fanni, E. Neri, *Introduction to Artificial Intelligence*. Springer, 2023.

J.L. Mezrich, *Is Artificial Intelligence (AI) a Pipe Dream? Why Legal Issues Present Significant Hurdles to AI Autonomy*. Am J Roentgenol, 2022.

E. Neri *et al.*, *Artificial intelligence: Who is responsible for the diagnosis?* La radiologia medica, 2020.

E. Neri *et al.*, *Explainable AI in radiology: a white paper of the Italian Society of Medical and Interventional Radiology*. La radiologia medica, 2023.

A.Y. Ng *et al.*, *Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer*. Nature Medicine, 2023, 3044 ss.

C. Privitera, *Intelligenza artificiale: cosa deve sapere il radiologo*. Giornale Italiano di Radiologia Medica, 2019.

S. Secinaro, *The role of artificial intelligence in healthcare: a structured literature review*. BMC Medical Informatics and Decision Making, 2021.

S. Triberti *et al.*, *A "Third Wheel" Effect in Health Decision Making Involving Artificial Entities: A Psychological Perspective*. Frontiers in Public Health, 2020.

Profili applicativi

NOZIONI INTRODUTTIVE:

Tra i sistemi di IA, i sistemi di *machine learning* (ML) sono quelli più diffusi, grazie alla loro capacità di apprendere da grandi quantità di dati. Benché molto spesso i due termini siano usati in modo interscambiabile, è importante chiarire che, se tutto ciò che riguarda il *machine learning* rientra nell'intelligenza artificiale, l'intelligenza artificiale non include solo il *machine learning*.

Esistono due principali categorie di modelli di *machine learning*: *machine learning* supervisionato e apprendimento non supervisionato. La differenza tra queste due tipi viene definita dal modo in cui ciascun algoritmo apprende i modelli per risolvere un *task*.

***Machine learning* supervisionato.** L'algoritmo assume come input un set di dati già etichettato e cerca di apprendere la relazione tra l'insieme delle *feature* che si stanno osservando e l'etichetta associata all'oggetto sotto osservazione. L'etichettatura del dato può derivare da fatti passati che hanno determinato la conoscenza dell'etichetta, oppure da un'etichettatura manuale fatta dall'uomo. Ad esempio, dato un insieme di radiografie, l'etichettatura può descrivere il fatto che si sta osservando la presenza di tumore oppure l'assenza di tumore. Esempi di modelli supervisionati sono modelli di classificazione come *Decision Tree*, Reti Neurali, *Super Vector Machine*, ecc.

***Machine learning* non supervisionato.** Il *machine learning* non supervisionato utilizza un approccio che cerca di estrarre dei pattern dai dati senza la guida di un'etichettatura esistente. Un esempio di approccio non supervisionato è il *clustering*, in cui, in assenza di etichette, si cercano di identificare nei dati gruppi di osservazioni simili sulla base delle caratteristiche che identificano l'oggetto sotto osservazione.

Tali modelli di *machine learning* trovano applicazione in contesti svariati: in medicina, in finanza, nelle pubbliche amministrazioni, nell'industria, ecc.

I modelli sia supervisionati che non supervisionati possono trattare dati di natura differente: dati tabulari, immagini, serie temporali, dati spazio-temporali, testi. Chiaramente, considerato il tipo di dato da trattare è necessario individuare il modello di *machine learning* più idoneo a sfruttare e trattare le dimensioni che caratterizzano il dato stesso. Ad esempio, per il trattamento di immagini e testo i modelli neurali sono la scelta migliore, mentre per dati di tipo tabulare modelli basati su alberi decisionali sono quelli più adeguati.

PROBLEMATICHE APPLICATIVE:

Tra le problematiche da considerare quando si decide di sfruttare sistemi di IA basati sul ML si possono sicuramente elencare:
- Qualità del dato che si vuole usare per l'apprendimento di tali modelli.
- Rappresentatività della popolazione di riferimento.

- Complessità e trasparenza dei modelli.
- Rischi legati alla violazione del diritto alla privacy.

Qualità del dato. Per poter apprendere un modello di *machine learning* è importante sottolineare la necessità di una collezione di dati e di un loro pre-processamento adeguato. I dati devono essere: *completi*, cioè rappresentare tutte le casistiche possibili; *aggiornati*, cioè che facciano riferimento a un periodo temporale adeguato al task; *accurati*, cioè per ogni *feature* osservata è necessario valutare la presenza di errori, valori mancanti, e di ogni altra anomalia.

Rappresentatività della popolazione di riferimento. Durante la preparazione del dato, prima dell'applicazione di un algoritmo per il *learning* di un modello un aspetto importante da valutare è la presenza di eventuali *bias* rispetto ad alcuni gruppi sottorappresentati e sbilanciamenti rispetto ad alcune classi.

Il *dato sbilanciato* potrebbe portare a delle performance del modello non adeguate, dunque a inaccuratezze importanti durante l'applicazione del modello. Esistono dei contesti in cui lo sbilanciamento è intrinseco al caso applicativo. Consideriamo, per esempio, un contesto applicativo come l'identificazione di transazioni finanziare fraudolente. Dato che le transazioni fraudolente sono molto meno rispetto a quelle non-fraudolente, un modello che cerchi di imparare a distinguere tra transazioni finanziare fraudolente e non con un *dataset* sbilanciato potrebbe presentare qualche difficoltà, a causa del limitato numero di esempi di azioni fraudolente. In presenza di tali situazioni è fondamentale un *pre-processing* che aiuti l'algoritmo a derivare un algoritmo che sopperisca a tale mancanza.

I *bias* nei dati spesso sono dovuti alla mancanza di diversità della coorte, che potrebbe dipendere, o non dipendere, da ragione tecniche.

Ad esempio, se consideriamo il contesto medico le ragioni potrebbero essere:
- **Criteri dello studio medico.** I dati sono stati raccolti per una certa popolazione, ad esempio per uno studio relativo alla sola popolazione adulta.
- **Progettazione non corretta del processo di raccolta del dato.** Ciò può accadere, generalmente, ogni volta che la popolazione utilizzata per addestrare, validare e testare il modello non rifletta la popolazione target del contesto clinico in cui il modello verrà applicato. Questa discrepanza potrebbe generare un'ampia gamma di pregiudizi, tra cui *bias temporali* (c'è un disallineamento tra il momento in cui il modello è stato sviluppato e il momento in cui il modello viene distribuito), *bias geografici* (il modello è stato sviluppato utilizzando solo i dati di un'area geografica), *spectrum bias* (la popolazione del set di dati utilizzata per sviluppare il modello non ha una rappresentazione reale dello spettro degli stati patologici – gravità, stadio ecc. – della popolazione target).
- **Uso secondario dei dati che sono stati raccolti per altri scopi**. Un esempio di tale pratica è l'uso dei codici ICD (classificazione internazionale delle malattie) per scopi di diagnosi predittiva nelle applicazioni DL. Questi codici erano originariamente destinati a scopi di fatturazione e potrebbero non descrivere correttamente il reale stato di salute del paziente.

Ragioni non tecniche di bias sono invece dovute alla storica omissione di alcune popolazioni dagli studi clinici, e al riflesso dei pregiudizi e della discriminazione

umana nel set di dati. Il requisito di *diversità, non discriminazione ed equità* prescrive che i dati di addestramento debbano includere tutte le categorie potenzialmente vulnerabili di individui rilevanti per lo studio. Numerosi esempi in letteratura mostrano come i pregiudizi discriminatori influenzino i risultati del *machine learning*. Ad esempio, Sayyed-Kalantari ha studiato il pregiudizio di una rete neurale convoluzionale profonda (CNN) all'avanguardia nell'assegnare la diagnosi corretta alle immagini radiografiche del torace. La CNN è stata addestrata su tre diversi grandi set di dati, e si è potuto così dimostrare come il tasso di sotto diagnosi fosse costantemente più alto per le donne e per i soggetti con un basso status socioeconomico.

Complessità e trasparenza dei modelli. Spesso, in contesti in cui il dato usato per l'addestramento è non tabulare, i modelli che presentano delle performance migliori sono modelli neurali, la cui complessità pone seri ostacoli alla comprensione del ragionamento alla base dell'algoritmo, da cui è derivata la decisione finale. Soprattutto in contesti critici come la medicina, capire il comportamento del modello è di fondamentale importanza per diversi aspetti: 1) per verificare il funzionamento del modello e poter eventualmente intervenire; 2) per identificare eventuali bias che possono essere nascosti da tale complessità; 3) per poter usare IA in modo consapevole e critico e non come un oracolo; 4) per poter capire il ragionamento e sfruttarlo per imparare dalla macchina, in modo da creare un ciclo virtuoso "la macchina impara dall'uomo e l'uomo dalla macchina".

Gli informatici stanno affrontando in diversi modi il problema della trasparenza: sviluppando modelli trasparenti *by design*, oppure cercando di spiegare i modelli complessi. Una recente linea di ricerca si pone anche l'obiettivo di rendere la spiegazione interattiva e adeguata alle capacità cognitive dell'utente. L'idea è lo sviluppo di modelli di spiegazione incrementali che, su richiesta dell'utente, possano dare vita a un *what-if* scenario, capace di fornirgli gli elementi giusti per una profonda comprensione del ragionamento. Chiaramente, la sfida maggiore, in questo caso, consiste nella idoneità del modello a fornire spiegazioni che si adeguino alle variabili capacità e necessità del singolo interlocutore umano. Utenti diversi potrebbero avere necessità completamente diverse.

La mancanza di trasparenza è il motivo principale della possibile scarsa fiducia dell'utente verso i sistemi di IA.

Rischi legati alla violazione del diritto alla privacy. Spesso i dati che si usano per l'addestramento dei modelli di *machine learning* sono dati personali, per cui la semplice psudonimizzazione potrebbe non essere sufficiente a garantire la protezione della privacy. In letteratura, purtroppo è stato dimostrato che, sebbene i dati non siano direttamente accessibili, è possibile sviluppare algoritmi sofisticati che riescano con una certa probabilità a inferire l'appartenenza a uno specifico individuo di un dato di training del modello di *machine learning*, semplicemente interrogando il modello stesso. È dunque necessario applicare tecniche di protezione che impediscano questo tipo di attacchi, quali la *differential* privacy, consapevoli però del fatto che il rumore inserito da tali meccanismi potrebbe ridurre l'efficacia dei modelli. In questo contesto, la sfida principale è, in definitiva, trovare il miglior compromesso tra protezione garantita e qualità del modello.

Quesiti

DA PARTE DEI GIURISTI:

1) *È possibile (e se sì, con quali modalità) sviluppare il dispositivo IA in modo da renderlo più trasparente sia con riferimento alla "logica generale" di funzionamento, che a quella "specifica" di risoluzione del determinato caso clinico?*

Nel campo dell'*Explainable* IA esistono diverse categorie di spiegazioni. Una prima distinzione importante riguarda i modelli di IA che offrono spiegazioni "dall'interno" (definiti *explainable by design*), ossia modelli che generano una spiegazione in congiunzione con la predizione. L'altra opzione è quella di creare modelli di spiegazione esterni (chiamati *post-hoc*), che aggiungono una spiegazione "accanto" alla predizione generata dall'IA. Ovviamente, la prima soluzione è ottimale, poiché la spiegazione è fornita direttamente con la predizione dal modello di IA. Tuttavia, i modelli *explainable by design* presentano ancora molti limiti tecnici: rendono più difficile ottenere buone performance del modello di IA, e le tempistiche per ottenere le spiegazioni sono molto lente. Pertanto, da qui in poi ci concentreremo solo sugli algoritmi di spiegazione *post-hoc*, in cui la spiegazione è fornita separatamente rispetto al modello di IA. Questi metodi consentono di utilizzare modelli di IA di qualsiasi tipo, anche estremamente complessi, cioè quelli cui siamo maggiormente interessati. Infatti, quando si lavora con una notevole quantità di dati, che vengono raccolti ogni giorno, un modello di IA semplice non riesce a ottenere buoni risultati in termini di correttezza delle predizioni.

Dato un modello di spiegazione *post-hoc*, possiamo avere vari tipi di spiegazioni. In particolare, possiamo distinguere le spiegazioni locali e globali, che nella domanda sono riferite, rispettivamente, come "spiegazioni della logica specifica" e "generali". Nel caso delle spiegazioni locali, il modello crea una spiegazione specifica per il dato in input. Ogni nuovo dato che viene passato ottiene una spiegazione diversa. Tecnicamente, questo tipo di modelli analizzano lo spazio attorno al dato in input per comprendere al meglio il ragionamento dell'IA nel caso specifico. Quindi, in ambito medico, la spiegazione può essere riferita a un determinato caso clinico. Per quanto riguarda le spiegazioni di tipo globale, l'obiettivo è spiegare in generale il comportamento del modello di IA. In pratica, la spiegazione riguarda il modello nel complesso, senza essere specifica su un particolare utente (o paziente, in questo caso). Di conseguenza, diversamente dalla spiegazione locale, questo tipo di spiegazioni viene generato una volta sola, dopo l'allenamento del modello di IA, e può essere consultato per comprendere i principi guida del nostro modello.

Dato un modello di IA, per ottenere una spiegazione della logica specifica e di quella generale, possono quindi essere estratte spiegazioni sia locali che globali.

2) *È possibile (e se sì, con quali modalità) consentire al dispositivo IA di confrontare il suo primo "output" con le valutazioni del medico e del paziente (ad es.: rifiuto della soluzione proposta, con richiesta di una nuova che tenga conto di determinati parametri, magari per evitare trattamenti che comportino una degenza di lungo periodo, etc.)?*

Negli ultimi anni, i ricercatori nel campo dell'IA stanno sempre più adottando un approccio "*user-centric*": l'idea è che, per migliorare i modelli di IA, sia necessario

considerare l'utente come parte integrante del processo. Un esempio motivazionale, in questo contesto, è quello di un modello di IA tradizionale applicato nell'ambito medico che, quando non è sicuro della propria decisione in rapporto a un paziente specifico, attualmente fornisce comunque una predizione. Tuttavia, questa scelta potrebbe non essere ottimale: ad esempio, se il medico si rende conto che il modello commette frequenti errori, potrebbe non considerarlo affidabile e smettere di utilizzarlo. Per superare questa limitazione, l'idea è quella di presentare al destinatario finale non solo il risultato, ma anche informazioni utili, come la confidenza del modello nella scelta proposta, così da consentire all'utente di decidere come agire con maggiore cognizione di causa, e persino di fornire suggerimenti al modello.

Questi argomenti, con varie sfaccettature, sono oggetto di ricerca in diversi campi: *learning to defer*, *learning to reject* e *interactive learning*. Tutti questi approcci hanno in comune il fatto che, se l'algoritmo è incerto riguardo alla predizione per un determinato record di input, il sistema si astiene dall'emettere una decisione. La definizione di incertezza del modello varia a seconda degli approcci considerati: ci sono sistemi che scartano dati che si discostano dalla distribuzione statistica appresa durante la fase di addestramento, altri che escludono dati "avversariali", cioè considerati innaturali o con valori anomali e forzati. Inoltre, una delle linee di ricerca più promettenti è quella che si astiene dalla predizione quando il dato in input è troppo vicino al limite decisionale del modello di IA. In tutti questi casi, quando il modello decide di astenersi dalla predizione, la decisione viene lasciata all'utente esperto che sta utilizzando il modello.

3) *È possibile (e se sì, con quali modalità) individuare la causa di un errore diagnostico della IA in modo sufficientemente chiaro da poter distinguere le ipotesi di malfunzionamento della IA da quelle di obiettiva difficoltà della diagnosi? È possibile (e se sì, con quali modalità) individuare la causa di un malfunzionamento della IA in modo sufficientemente chiaro da poter distinguere le ipotesi di difetto progettuale da quelle di errore "fisiologico"? È possibile (e se sì, con quali modalità) individuare la causa di un malfunzionamento della IA in modo sufficientemente chiaro da poter capire all'opera di quale fra i vari soggetti che partecipano alla fase di produzione può essere addebitata?*

Quando si lavora con modelli di IA vi è una procedura di buona norma da seguire, che attiene alle fasi di sviluppo di un modello. Possiamo identificare 4 punti fondamentali:

1. identificazione del dato da utilizzare per allenare il nostro algoritmo: qual è l'obbiettivo, che tipo di dato abbiamo (reale, sintetico), quali sono le dimensioni di questo dato;
2. *data cleaning* e preparazione: una volta scelto il dato da usare, bisogna pulirlo ed organizzarlo al meglio per poterlo usare. Esempi di procedure da applicare in questa fase sono l'analisi dei *missing values*, cioè quei *records* che hanno dei valori mancanti per qualche variabile (ad esempio potrebbe essere mancante l'età di un paziente, oppure il sesso). In questo caso ci sono varie tecniche per gestire i *missing values*, che possono essere divise in due gruppi: l'eliminazione, della riga o della colonna in analisi, oppure la sostituzione del valore mancante con un valore fittizio;

3. scelta del modello di AI: al giorno d'oggi ci sono molti modelli di IA che possono essere utilizzati, come *Random Forest*, *Neural Networks* etc. In dipendenza delle caratteristiche del dato in input, bisogna scegliere il modello migliore;
4. validazione e *testing*: dopo aver allenato il modello, bisogna capire se esso funziona in modo appropriato. In particolare, ci sono varie tecniche per validare la bontà del modello ottenuto. Come prima cosa, il dato in input viene diviso in *train* e *test*, in modo da poter allenare su una parte di dato e validare su un dato che il modello non ha mai visto. Inoltre, è buona norma effettuare *cross-validation*: provare ad allenare il modello più volte, su diverse divisioni di *train* e *test*, in modo da ottenere un modello robusto.

A questo processo, possiamo anche aggiungere la parte di spiegazione, che potrebbe essere in grado di individuare gli elementi che hanno determinato la predizione (o, nel caso medico, la diagnosi). In particolare, con l'aiuto di utenti esperti nel settore in cui il modello di IA lavora, si possono individuare i malfunzionamenti in ragione delle spiegazioni fornite. In questo ambito si riscontrano alcuni lavori preliminari, che si inseriscono nel campo del *learning to reject* e del *learning to defer*.

Tuttavia, offrire una spiegazione all'utente finale, esperto nel proprio settore, è indubbiamente importante, ma in alcuni casi potrebbe non essere sufficiente per individuare i malfunzionamenti del modello. Pertanto, per una validazione accurata da un punto di vista tecnico, una possibilità è utilizzare i *white box explanation methods*. Questi metodi forniscono spiegazioni *post-hoc* che sfruttano le conoscenze strutturali del modello di IA in analisi. Ad esempio, un metodo *white box* potrebbe essere specifico per le *Neural Networks*, in particolare per le *Convolutional Neural Networks*. La conoscenza strutturale del modello di IA consente a queste spiegazioni di essere molto specifiche, focalizzate sul funzionamento del modello. Ne derivano spiegazioni più complesse, difficili da comprendere per un utente non esperto di informatica, ma che possono essere ideali per eseguire operazioni di debug, ovvero comprendere il funzionamento interno del modello e, di conseguenza, identificare possibili malfunzionamenti.

4) *È possibile (e se sì, con quali modalità) simulare l'esito della diagnosi di una IA in un dato quadro clinico, tenendo presente il livello di sviluppo tecnologico della IA al momento dei fatti?*

Quando parliamo di modelli di IA "tradizionali", in cui abbiamo una fase di analisi del dato, apprendimento da parte del modello e validazione dei risultati ottenuti, è sempre possibile simulare l'esito di una diagnosi data dal modello. Questo perché, esattamente come per tutti i sistemi informatici in vigore (anche senza l'uso di IA), bisogna mantenere traccia del modello pubblicato, per poter effettuare anche in un lontano futuro analisi di responsabilità o de-bug (*e.g.* sistemazione di piccoli errori che possono essere scoperti durante l'utilizzo del sistema informatico).

La situazione potrebbe sembrare più complicata quando si parla di modelli di IA con apprendimento continuo, ma in realtà non lo è. Con il termine apprendimento continuo si intendono modelli di IA in grado di apprendere continuamente dagli input che ricevono, con l'obbiettivo di migliorare le loro

performance. In questo caso, nonostante l'algoritmo sia in grado di apprendere continuamente, per poterlo utilizzare è necessario "estrarre" una copia del modello di IA in un dato momento, validarne la correttezza e renderlo pubblico per l'uso. Come esempio si può pensare al famoso ChatGPT: un algoritmo di apprendimento continuo di cui possiamo usare diverse versioni (la n. 2, 3 o 4), che sono degli "estratti" dell'apprendimento continuo. Il risultato è, quindi, una versione del modello di IA, estratta dal modello che continua ad imparare, generata con un apprendimento continuo, ma che si fa immodificabile nel momento in cui diventa stabile. Di conseguenza, questa versione stabile può essere usata e mantenuta anche per simulazioni future, nonostante il modello principale continui ad imparare.

5) *Quali accorgimenti sono necessari per assicurarsi che gli output resi da un sistema in apprendimento continuo mantengano la medesima qualità e non si deteriorino nel tempo?*

Nel campo dei modelli di IA di apprendimento continuo, l'idea è quella di avere modelli in grado di apprendere continuamente dagli input ricevuti. In generale, gli algoritmi ad apprendimento continuo, per poter essere pubblicati e, di conseguenza, usati, devono seguire lo stesso iter di cui già si è detto rispondendo alla domanda n. 3: ogni volta che un modello viene creato, devono essere seguiti i 4 punti fondamentali di identificazione ed analisi del dato, pulizia del dato, scelta del modello migliore nel contesto in analisi e validazione dei risultati ottenuti.

Rispetto a un modello ad apprendimento continuo, la procedura richiede, tuttavia, un ulteriore passaggio: nonostante il modello in questione sia in grado di imparare costantemente, per poterlo pubblicare è necessario "fermarlo" temporaneamente e, quindi, estrarne una copia, che deve poi essere validata prima di essere effettivamente utilizzata. In pratica, rispetto alla procedura descritta in rapporto alla domanda n. 4, vi è di diverso che il modello principale continua ad essere allenato, mentre la versione estratta è stata "fermata".

Rispettando questi passaggi, non può darsi un deterioramento delle qualità degli output del modello, perché, in tal caso, il modello non passerebbe i controlli di correttezza e non verrebbe pubblicato per l'utilizzo.

6) *A fronte della possibilità di revocare il consenso al trattamento dei propri dati personali, con conseguente necessità di rimuoverli* pro-futuro *dai sistemi di utilizzo, una IA ha effettivamente la possibilità di rimuovere un dato dal proprio processo valutativo e, dunque, può "disimparare" quanto appreso mediante quel dato?*

Recentemente, con l'avvento dell'IA generativa e dell'AI Act, si è iniziato a discutere anche del tema del "*forgetting*". Attualmente, la soluzione per ottenere il *forgetting* di un dato *record* consiste nell'eliminare quel record dal *dataset* e ri-allenare il modello di IA da capo. Tuttavia, questo approccio presenta diversi limiti, tra cui tempi molto lunghi e un notevole dispendio di risorse energetiche. Di conseguenza, molte aziende esitano a eliminare i propri dati, anche in ragione dei costi.

Per superare tali problemi tecnici, sono state proposte diverse metodologie. Una delle prime proposte risale al 2019, e ha come obiettivo l'alterazione dei pesi di un modello di IA (in questo caso *Neural Networks*) in modo tale che le informazioni relative ai record da eliminare vengano dimenticate. Inoltre, questo

metodo richiede solo la conoscenza dei record da eliminare e non di tutto il *dataset* su cui il modello è stato addestrato.

Attualmente c'è un grande interesse per l'*unlearning*, tanto che persino Google ha lanciato una sfida per lo sviluppo di tali algoritmi. Tuttavia, la ricerca in questo ambito è ancora in fase embrionale e presenta diversi limiti. Innanzitutto, i modelli sottoposti all'*unlearning* non mantengono le stesse prestazioni predittive di prima, perdendo quindi in precisione. Inoltre, i metodi di *unlearning* sono ancora piuttosto lenti, pur impiegando meno tempo rispetto al ri-addestramento completo.

DA PARTE DEI MEDICI:

7) *Fino a che punto può spingersi l'informatica nel tentativo di rendere "spiegabile" l'IA? A che punto siamo di questo percorso?*

Molti ricercatori stanno lavorando in questo contesto e già disponiamo di diversi algoritmi di spiegazione utilizzabili. In quest'ambito sono da considerare tantissime variabili: prima fra tutte, la tipologia di spiegazione (regole, controfattuale, *feature importance*), che è dipendente anche dalla tipologia di dato in input (per dati tabulari abbiamo spiegazioni diverse che non per le immagini). Purtroppo, però, la ricerca in materia di *Explainable IA* è ancora giovane e tante sono le sfide aperte. Come prima cosa, i metodi al momento disponibili non sono generici: molti funzionano solo per alcuni tipi di dato specifici, quindi richiedono ulteriori analisi quando vengono applicati a contesti nuovi. Inoltre, la scarsa maturazione del settore comporta difficoltà nella valutazione della qualità e della correttezza della spiegazione ottenuta. Mancano, quindi, validazioni obbiettive e misure quantitative. Una possibile soluzione, su cui molti ricercatori stanno lavorando, è l'uso di esperti per validare la bontà della spiegazione. Oltre a questi problemi, vi è anche quello della robustezza del modello proposto: spesso i metodi per generare spiegazioni rispondono a tecniche di randomizzazione, oppure risentono di approssimazioni per poter risolvere problemi altrimenti intrattabili. Queste tecniche, per quanto usate in ambito informatico e matematico, hanno come limite il fatto che, dato un esempio in input per due volte consecutive, si potrebbero ottenere due risposte differenti.

8) *Alla luce di ciò che manca oggi e di ciò che sarà raggiungibile nell'arco di qualche anno, in che misura la classe medica dovrebbe essere formata da un punto di vista informatico per una corretta comprensione, e conseguentemente un corretto utilizzo, di questi strumenti?*

Come già descritto più volte in questo documento, XAI è un ambito ancora molto giovane. Di conseguenza, allo stato attuale molte spiegazioni che vengono generate non sono comprensibili per un utente che non possieda conoscenze di informatica. Ad ogni modo, più che aggravare la formazione dei medici, riteniamo opportuno investire, e proseguire, nella ricerca sui modelli informatici di XAI, con l'obbiettivo di rendere le spiegazioni interpretabili anche da utilizzatori non dotati di competenze informatiche particolarmente elevate. Sarà comunque necessario, per il medico, disporre di un bagaglio minimo di nozioni informatiche, ma non crediamo necessaria una specializzazione in tal senso. Tale preferenza deriva dalla consapevolezza del contesto più ampio in cui la XAI virtualmente si colloca, che

comprende non solo l'ambito medico, ma anche quello legale, commerciale, educativo etc.: non si può immaginare che professionisti dei più diversi settori debbano tutti specializzarsi in informatica allo scopo di comprendere spiegazioni altrimenti, per loro, indecifrabili.

Attualmente, uno degli ambiti di ricerca più interessanti è quello della validazione delle spiegazioni con un approccio di co-design: tramite la cooperazione tra utenti e modelli, si studia come elaborare forme di spiegazione che sposino le aspettative dei destinatari e che siano adeguate alle loro conoscenze ed esperienze. Già numerosi sono i contributi scientifici a questo riguardo, anche specifici per l'ambito medico.

9) *Maggiore complessità significa, solitamente, maggiore vulnerabilità. Come garantire la sicurezza dei big data dei pazienti, evitando episodi di perdita, dispersione o corruzione dei dati? Si può garantire che un algoritmo di pseudonimizzazione dei dati renda effettivamente impossibile risalire all'identità di un paziente, come se essi fossero di fatto anonimizzati?*

In ambito privacy riscontriamo due principali linee di indagine: la privacy sul dato e la privacy sul modello di IA. Nel primo caso, il focus è sul dato in input, che, anche se "pseudonimizzato", poterebbe comunque essere re-identificato grazie ad altre informazioni contenute nel dato che abilitano una identificazione indiretta (quasi-identificatori). Nel secondo caso, invece, assumiamo che il dato sia privato, visibile solo al proprietario del modello di IA. In questa seconda ipotesi, quindi, viene valutata la privacy del modello di IA, cioè la sua capacità di proteggere le informazioni "imparate" durante la fase di allenamento.

In entrambi i casi, il processo segue le direttive della GDPR: una prima fase di valutazione del rischio di privacy, sia sul dato che sul modello, seguita da una fase di protezione del rischio di privacy, fatta in dipendenza dai risultati ottenuti nella prima fase.

Bibliografia

Ian Goodfellow, Yoshua Bengio, Aaron Courville, *Deep Learning*. MIT Press, 2016.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, Dino Pedreschi, *A Survey of Methods for Explaining Black Box Models*. ACM Comput. Surv., 51(5): 93, 2019, 1 ss.

Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, Salvatore Rinzivillo, *Benchmarking and survey of explanation methods for black box models*. Data Min. Knowl. Discov., 37(5), 2023, 1719 ss.

Andreas Theissler, Francesco Spinnato, Udo Schlegel, Riccardo Guidotti, *Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions*. IEEE Access 10, 2022, 100700 ss.

M. Abadi, A. Chu, I.J. Goodfellow, H.B. McMahan, I. Mironov, K. Talwar, L. Zhang, *Deep learning with differential privacy*. ACM Conference on Computer and Communications Security, 2016, 308 ss.

R. Shokriand, V. Shmatikov, *Privacy preserving deep learning*. ACM Conference on Computer and Communications Security, 2015, 1310 ss.

L. Zhao, Q. Wang, Q. Zou, Y. Zhang, and Y. Chen, *Privacy-preserving collaborative deep learning with unreliable participants*. IEEE Transactions on Information Forensics and Security, 15, 2020, 1486 ss.

S. H. Park and K. Han, *Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction*. Radiology, 286(3), 2018, 800 ss.

A. Casey, B. S. Schwartz, W. F. Stewart, and N. E. Adler, *Using electronic health records for population health research: a review of methods and applications*. Annual review of public health., 37, 2016, 61 ss.

Yadav, M. Steinbach, V. Kumar, and G. Simon, *Mining electronic health records (ehrs): a survey*. ACM Computing Surveys (CSUR), 50(6), 2018, 85 ss.

Cecilia Panigutti, Alan Perotti, André Panisson, Paolo Bajardi, Dino Pedreschi, *FairLens: Auditing black-box clinical decision support systems*. Information Processing & Management, 58(5), 2021, 102657.

F. Doshi-Velez, B. Kim, *Towards a rigorous science of interpretable machine learning*. arXiv preprint arXiv:1702.08608, 2017.

L. Seyyed Kalantari, G. Liu, M. McDermott, and M. Ghassemi, *Chexclusion: Fairness gaps in deep chest x-ray classifiers*. arXiv:2003.00827, 2020.

Davide Bacciu, Paulo J G Lisboa, Alfredo Vellido, *Deep Learning in Biology and Medicine*. World Scientific Publishing, 2022.

Clara Punzi, Roberto Pellungrini, Mattia Setzu, Fosca Giannotti, Dino Pedreschi *AI, Meet Human: Learning Paradigms for Hybrid Decision Making Systems*, arXiv 2024.

Francesca Naretto, Anna Monreale, Fosca Giannotti *Evaluating the privacy exposure of interpretable global explainers*. 2022 IEE 4th Interantional Conference on Cognitive Machine Intelligence, 2022.

Aditya Golatkar, Alessandro Achille, Stefano Soatto *Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks*, arVix, 2020.

Pratibha Kumari, Joohi Chauhan, Afshin Bozorgpour, Boqiang Huang, Reza Azad, Dorit Merhof *Continual Learning in Medical Image Analysis: A Comprehensive Review of recent Advancements and Future Prospects*, arXiv 2023.

RILIEVI CONCLUSIVI

Verso l'implementazione di un modello applicativo

I MOMENTI DELL'ATTO MEDICO RADIOLOGICO:

Secondo il documento della Società Italiana di Radiologia Medica e Interventistica (SIRM) pubblicato nel 2019, l'atto medico radiologico, inteso come prestazione professionale specialistica, «*ha finalità diagnostiche e/o interventistiche e consta di una serie di momenti strettamente interdipendenti ed inscindibili*».

I momenti in questione possono essere schematizzati come segue:

1. valutazione della richiesta di prestazione del medico prescrivente;
2. inquadramento clinico-laboratoristico-anamnestico;
3. giustificazione dell'esame proposto;
4. raccolta del consenso all'atto medico;
5. identificazione, ottimizzazione ed esecuzione dell'esame;
6. documentazione iconografica;
7. interpretazione/refertazione/comunicazione/discussione con il clinico;
8. archiviazione dell'esame.

L'insieme di questi momenti costituisce un *workflow*, che, ovviamente, non è sempre esente da errori. Negli ultimi anni l'intelligenza artificiale è stata proposta come potenziale soluzione a tutti i problemi che possono insorgere in ciascuno dei diversi momenti di questo *workflow*, dalla valutazione dell'appropriatezza della richiesta fino all'ottimizzazione della dose. Tuttavia, è necessario precisare che non sempre la stessa attenzione è stata dedicata ai diversi step e, più specificamente, il "task interpretativo", per quanto rappresenti solo una parte di tutto l'iter, ha riscosso un maggiore interesse da parte del pubblico, generalista o meno, con conseguente maggiore investimento da parte delle aziende.

Per tale ragione, volendo esemplificare l'integrazione di un algoritmo di IA nella pratica clinica, quello "interpretativo" rappresenta il task ideale. Più nello specifico, si descriverà l'integrazione di un software di IA dedicato alla interpretazione di immagini radiografiche per la *detection* di fratture in pazienti traumatizzati.

UN ESEMPIO DI INTEGRAZIONE DELL'IA NELLA PRATICA CLINICA:

Rappresentando ad oggi un campo relativamente nuovo, non esistono, allo stato attuale, dei veri e propri standard di integrazione di questi software. Pertanto, nella descrizione di questo *workflow*, ci si troverà a introdurre alcune possibili variabili.

Certamente, il software si troverà a dialogare con i sistemi informatici standardizzati di cui attualmente ogni radiologia è provvista, il RIS e il PACS. Il RIS, acronimo di Sistema Informativo Radiologico (in inglese *Radiology Information System*), è un sistema utilizzato per gestire vari aspetti amministrativi legati ai pazienti, come la prenotazione delle visite mediche, l'accettazione, la refertazione e la firma digitale. D'altra parte, il PACS, ovvero Sistema di

Archiviazione e Trasmissione di Immagini (in inglese *Picture Archiving and Communication System*), è utilizzato per archiviare, trasmettere, visualizzare e stampare immagini diagnostiche digitali. Grazie al PACS, una struttura sanitaria può archiviare le immagini su supporti digitali come CD, DVD. Una corretta integrazione di un software di IA non può prescindere da un dialogo efficace con questi due sistemi. Per fare un esempio, il software deve essere in grado di processare le immagini nello stesso formato in cui sono archiviate nel PACS, ovverosia il formato c.d. DICOM.

Una volta messo in condizione il software di "comunicare" con i sistemi informatici, e quindi con la *workstation* del radiologo, è necessario stabilire i tempi e le modalità di questa comunicazione. In proposito, possono essere individuati due differenti modelli di integrazione:

1. Le immagini, acquisite dal tecnico sanitario di radiologia medica, vengono inviate al PACS e, simultaneamente, al software di IA, che potrà essere installato localmente o in un cloud. Tutte le immagini, di tutti i pazienti, vengono pertanto processate dal software. In effetti, diversi software di *post-processing* vengono comunemente adottati in qualsiasi percorso diagnostico di *imaging* (e.g., software di ricostruzione tridimensionale di immagini di colonscopia virtuale), di cui il paziente non viene specificatamente informato. A questo punto, il software processerà le immagini e potrà restituire il suo output. L'output potrà essere fornito direttamente sul RIS (frattura si/frattura no), in modo da dare priorità di refertazione a esami positivi piuttosto che a esami negativi, o sul PACS, sotto forma di immagine variabilmente etichettata. Il software potrebbe individuare la frattura/le fratture sull'immagine, esprimendo un vero e proprio grado di confidenza diagnostica per ciascuna di queste fratture. Il radiologo, nel momento in cui visualizzerà la cartella radiologica del paziente, con tutte le immagini acquisite, dovrà visualizzare anche le immagini etichettate dal software di IA e integrarle nel proprio flusso decisionale per arrivare a una conclusione diagnostica. Il referto, così come le immagini, sotto forma di CD/DVD, verranno quindi fornite al paziente, includendo anche le immagini processate dal software.

In questo tipo di modello, tutte le immagini vengono processate e vengono restituite come output variabilmente al RIS e/o al PACS. Il radiologo, come tutte le informazioni di cui è in possesso prima di formulare la sua diagnosi, dovrà necessariamente tenerne di conto, e, ancora una volta, dovrà farlo per tutti i pazienti.

2. Le immagini, acquisite dal tecnico sanitario di radiologia medica, vengono inviate al PACS. Il radiologo visualizza la cartella radiologica del paziente con tutte le immagini acquisite e, laddove lo riterrà necessario sulla base dello specifico caso e del proprio expertise, anche previa eventuale acquisizione del consenso informato da parte del paziente, potrà inviare le immagini al software di IA installato localmente o nel cloud. Il software restituirà quindi l'output, che potrà essere o meno salvato sul PACS. Sulla base di tutte le informazioni in possesso, il radiologo formulerà la sua diagnosi. Il referto, così come le immagini, sotto forma di CD/DVD, verranno quindi fornite al paziente.

Un'integrazione così strutturata introdurrebbe una variabile "personale". È il radiologo che stabilisce quando e come utilizzare il software, il cui output verrà o

meno fornito anche al paziente insieme alle immagini e al referto. Il radiologo, più o meno esperto, potrà avere minore/maggiore fiducia nei confronti del software e stabilire arbitrariamente quando e come utilizzarlo, anche sulla base delle caratteristiche tecniche del software medesimo.

DAL PARTICOLARE AL GENERALE, DUE MODELLI APPLICATIVI A CONFRONTO:

L'assenza di sufficienti dati clinici e la varietà delle possibili applicazioni radiologiche impediscono di proporre degli standard di integrazione condivisa dell'IA in questa materia, donde la necessità di simularne l'applicazione solo con riguardo ad una specifica ipotesi diagnostica, addotta appunto a titolo di esempio; tuttavia, nonostante la sua inevitabile parzialità, l'esempio appena proposto ha illuminato una tensione di fondo nell'impiego dell'IA in ambito medico-diagnostico da cui si potrebbero trarre delle conclusioni di carattere generale.

In particolare, sembra possibile individuare nei modelli di integrazione sopra descritti due approcci alla questione che divergono da un punto di vista – verrebbe da dire – quasi ideologico, definibili il primo come di tipo "**efficientista**", il secondo come di tipo "**personalizzante**".

In una versione estrema di tali due modelli, **secondo quello "efficientista", l'IA non avrebbe ragione di essere distinta da un qualunque altro strumento clinico in dotazione al medico**, potendosi paragonare al bisturi per un chirurgo: il chirurgo neppure si pone il problema di rifiutare l'utilizzo dei bisturi fornitigli dalla struttura medica nel quale è collocato, né è necessario che avverta il paziente circa l'utilizzo di questo o quel modello di bisturi; lo stesso dicasi per l'IA: il radiologo la troverà già implementata nel suo *workflow* non potendosi esimere dall'utilizzarla, così come il paziente non dovrà essere avvertito rispetto al suo impiego, posto che esso sarebbe connaturato all'attività radiologica, esattamente come l'uso del bisturi è connaturato a quella chirurgica.

Il secondo modello, viceversa, sottende una concezione dell'IA alla stregua di uno strumento di assoluta eccezionalità, che imporrebbe al medico curante di utilizzarla in modo del tutto discrezionale, come se fosse un rimedio straordinario, il cui impiego, pertanto, potrebbe passare anche dal consenso informato del paziente.

È evidente come **entrambi questi modelli – nelle loro versioni estreme – presentino delle criticità**: il primo trascura eccessivamente, sino a banalizzarlo, l'impatto dell'IA in ambito diagnostico, sottraendo ai clinici qualsiasi gestione della stessa, mentre il secondo – all'opposto – esagera tale impatto e, con una visione quasi primitivistica, rifiuta di razionalizzarne l'impiego mediante regole standardizzate.

È chiaro, dunque, che una sfida cui sarà chiamata a rispondere la futura regolamentazione della materia sarà quella di **adottare un approccio equilibrato tra i due modelli appena proposti, che sappia cogliere l'eccezionalità dell'IA senza, tuttavia, rinunciare ad inserire anche tale strumento all'interno di procedure uniformi volte a garantire la salute dei pazienti.**

Ebbene, per chiudere il discorso, è opportuno fornire un esempio pratico di tale approccio equilibrato tra i due modelli proprio con riguardo all'ipotesi di *workflow* poc'anzi immaginata in relazione all'integrazione dell'IA nell'interpretazione di immagini radiografiche per la *detection* di fratture in pazienti traumatizzati.

Nello specifico, si potrebbe proporre un correttivo "personalizzante" rispetto al primo modello proposto, quello, cioè, di tipo "efficientista"; quest'ultimo, infatti, parrebbe più accettabile se, prima del salvataggio delle immagini sul PACS, fosse consentito al radiologo un vaglio ragionato sulle immagini stesse, affidandogli l'onere della decisione se accettare o rifiutare l'output dell'IA. In caso di accettazione dell'output la procedura continuerebbe come descritta, mentre, qualora l'output venisse rifiutato, l'immagine non verrà memorizzata sul PACS e non verrà fornita al paziente insieme al referto radiologico.

In tal modo, pur all'interno di una procedura regolamentata, non si priverebbe il radiologo di qualsiasi potere gestorio dell'IA, garantendogli di controllare tale strumento e non di esserne controllato.

Pisa, 9 novembre 2024

La sezione penalistica è stata curata da:	**La sezione civilistica è stata curata da:**	**La sezione medica è stata curata da:**	**La sezione informatica è stata curata da:**
Nicolò Amore	Andrea Cioni	Lorenzo Faggioni	Claudio Giovannoni
Marcello Sestieri	Diletta Corti	Salvatore C. Fanni	Anna Monreale
Antonio Vallini	Maria E. Lippi	Emanuele Neri	Francesca Naretto