Annali di studi religiosi
ISBN 979-12-218-2156-7
DOI 10.53136/979122182156731
pp. 559-577 (ottobre 2025)

INTELLIGENZA ARTIFICIALE E INGIUSTIZIA EPISTEMICA NELLA CURA E NELLA RICERCA DELLA MALATTIA MENTALE: UNO STATO DELL'ARTE

Eugenia Lancellotta

Abstract: Artificial Intelligence (AI) is becoming the focus of media attention and economic investments in many of the world's economies. However, the ethical and social issues raised by these new technologies are still in the early stages of exploration. In this article, I provide a critical overview of the scientific and philosophical literature on one of these issues: the relationship between AI and epistemic injustice in the field of mental health research and care. In psychiatry, epistemic injustice occurs when the epistemic contributions of individuals suffering from mental illnesses are unfairly discredited by doctors and researchers due to biases related to the identity of the patients, who are considered irrational and unreliable witnesses. While most of the scientific literature on this topic argues that the use of AI in psychiatry could exacerbate and reinforce the existing epistemic injustice, some voices highlight the potential opportunities offered by new technologies in the fight against this insidious form of oppression.

L'Intelligenza Artificiale (IA) si sta ponendo al centro dell'attenzione mediatica e degli investimenti economici di gran parte delle economie mondiali. Tuttavia, le questioni etiche e sociali che queste nuove tecnologie sollevano sono ancora in una fase primitiva di esplorazione. In questo articolo, fornisco una panoramica critica della letteratura scientifica e filosofica su una di queste questioni, ovvero il rapporto tra IA e ingiustizia epistemica nell'ambito della ricerca e della cura psichiatrica. In campo psichiatrico, l'ingiustizia epistemica si manifesta quando i contributi epistemici delle persone che soffrono di malattie mentali vengono ingiustamente screditati da parte di medici e ricercatori a causa di pregiudizi legati all'identità dei malati, ritenuti testimoni irrazionali e poco affidabili. Mentre la maggior parte della letteratura scientifica in merito sostiene che l'uso dell'IA in psichiatria potrebbe aggravare e rinforzare l'ingiustizia epistemica già esistente, alcune voci sottolineano le possibili opportunità aperte dalle nuove tecnologie nella lotta contro questa subdola forma di oppressione.

Keywords. Artificial Intelligence, Epistemic injustice, Psychiatry, Epistemic conformism, ADHD

Parole chiave: Intelligenza Artificiale, Ingiustizia epistemica, Psichiatria, Conformismo epistemico, ADHD

1. Introduzione

Il termine Intelligenza Artificiale (IA) fu coniato da John McCarthy nel 1956, che la definì come la scienza e la progettazione di macchine intelligenti (Frana e Klein 2021, p. 105). Nonostante si stia assistendo a un boom mediatico e di investimenti in tale tecnologia, il suo utilizzo nella pratica medica è ancora fortemente limitato (Monteith, Glenn, Geddes *et al.* 2022). Ciò si deve a diversi fattori, il primo dei quali è riconducibile al cosiddetto *paradosso produttivo*, concetto teorizzato dal premio Nobel per l'economia Robert Solow nel 1987 (Solow 1987). Secondo il paradosso, esisterebbe un *gap* di decenni tra la creazione e l'implementazione produttiva di una nuova tecnologia. La pratica medica non farebbe eccezione, anzi: un'applicazione dell'IA in quest'ambito risulta ulteriormente complicata dall'entità dei rischi che il funzionamento subottimale di queste nuove tecnologie potrebbe comportare.

Per apprezzare lo stato di (im)maturità attuale dell'IA nella pratica medica, è importante far riferimento alla *technology readiness level (TRL) scale*, una scala a 9 livelli inventata dalla NASA negli anni Settanta e tuttora in uso per misurare il livello di maturità di una tecnologia (Mankins 1995). Recentemente adattata per la valutazione dell'IA in ambito clinico (Fleuren, Thoral, Shillan *et al.* 2020), i livelli fino al 4 si riferiscono all'uso di prototipi di IA in ambito sperimentale, mentre i restanti livelli indicano che la tecnologia in questione ha superato lo stadio del prototipo, e che è stata anche impiegata su una popolazione più ampia di quella con cui è stato testato il prototipo. In recenti studi condotti sull'uso dell'IA in medicina intensiva (Fleuren, Thoral, Shillan *et al.* 2020; Van de Sande, van Genderen, Huiskens *et al.* 2021), il 90% circa di queste tecnologie si fermava al livello 5 della scala TRL, con la maggior parte di esse attestantesi al livello 4 o inferiori: nessuna tecnologia raggiungeva il livello 9 o era routinariamente inclusa nella pratica

medica. Inoltre, in un recente sondaggio condotto tra alcuni radiologi dell'American College of Radiologists (Allen, Agarwal, Coombs *et al.* 2021) — la radiologia è l'ambito medico in cui l'uso dell'IA è maggiormente diffuso — il 94% degli intervistati ha valutato la performance dell'IA come scarsamente affidabile.

L'IA si trova a uno stadio di maturazione bivalente nell'ambito della salute mentale. Da un lato, i chatbots psicoterapeutici hanno conosciuto un'esplosione nell'industria del benessere degli ultimi anni, complici i costi esorbitanti dei professionisti per la salute mentale nei paesi a sistema sanitario privato e le lunghissime liste d'attesa nei paesi a sistema sanitario pubblico o semi–pubblico (McAllen 2024). Nonostante i chatbots per il benessere psicologico presentino diversi benefici in termini di costi e accessibilità, non essendo controllati da un'autorità riconosciuta che ne garantisca la sicurezza e l'efficacia, essi presentano diversi rischi, tra cui, ad esempio, misinformazione e violazione della privacy dei dati degli utenti. Dall'altro, l'IA è ancora scarsamente usata in ambito clinico-psichiatrico. Il sistema sanitario nazionale inglese, NHS, ha solo di recente introdotto una app di IA, *Limbic* (Khaku 2023), per fare uno screening iniziale di chi richiede assistenza per la salute mentale; così la FDA, l'agenzia regolatrice del sistema sanitario nazionale americano, ha da poco approvato l'utilizzo su prescrizione medica di un chatbot per la depressione maggiore, *Rejoyn* (Cheng 2024).

Questo dato può risultare sorprendente se si pensa che è da più di mezzo secolo che i ricercatori tentano di utilizzare l'IA nell'ambito della salute mentale. Nel 1966 Joseph Weizenbaum creò Eliza, il primo prototipo di chatbot psicoterapeuta: molte delle idee alla base di questo programma sono ancora utilizzate nelle applicazioni di IA per la salute mentale di oggi (Berry 2018; Weizenbaum 1966). Tuttavia, la scarsa diffusione dell'IA in ambito psichiatrico può essere spiegata dalle seguenti limitazioni (Monteith, Glenn, Geddes *et al.* 2022). Innanzitutto, l'IA necessita di tanti dati di qualità per diagnosticare, distinguere e curare correttamente i disturbi mentali. Tuttavia, vi è una grande carenza di tali dati in ambito psichiatrico. I registri elettronici, da cui questi dati vengono in buona parte tratti, presentano diverse anomalie, perché non sono stati inizialmente pensati per scopi diagnostici o di ricerca. Per questo motivo, si registrano, tra le varie anomalie, discrepanze e diagnosi mancate

del Disturbo Post Traumatico da Stress; una mancata documentazione di idee o tentativi suicidali; una stima al ribasso di sintomi a cui è associato un forte stigma sociale; la mancata esplicitazione di una diagnosi per diversi pazienti a cui sono stati prescritti farmaci psicotropi; cambiamenti di codici linguistici nel corso del tempo. Alla scarsità di dati di qualità con cui alimentare i sistemi di IA, si aggiungono le complessità legate all'utilizzo degli algoritmi. Ad esempio, un aumento della diversità della popolazione in termini di età, sesso e immagini cerebrali ha contribuito alla riduzione della precisione predittiva di modelli di IA per il neuroimaging, con effetti sulla diagnosi di disturbi dello spettro autistico. La scarsa educazione tecnologica dei medici, che li porterebbe a un atteggiamento di deferenza eccessiva nei confronti dell'IA al cospetto di una decisione clinica da prendere, è un'altra importante limitazione nell'applicazione dell'IA in ambito psichiatrico, come lo è il rischio che i clinici e i lavoratori nel settore della salute mentale subiscano un processo di dequalificazione a causa dell'eccessiva dipendenza da tali sistemi. Infine, non va dimenticata l'attuale scarsità di misure di validazione per l'IA in ambito clinico, dovuta in parte alle precedenti limitazioni di questi sistemi e in parte al fatto che le regolamentazioni di tecnologie mediche tradizionali non sono state pensate per essere applicate all'IA.

In questo articolo, condurrò un breve stato dell'arte su un aspetto particolare e ancora poco studiato dell'IA nella ricerca e nella cura della malattia mentale, ovvero la relazione tra IA e ingiustizia epistemica. L'ingiustizia epistemica è una forma di danno epistemico particolarmente diffuso nella ricerca e soprattutto nella cura della malattia mentale, e la ricerca filosofica sul tema si è fatta particolarmente vivace negli ultimi anni (Crichton, Carel e Kidd 2017; Kidd, Spencer e Carel 2025). La nozione di ingiustizia epistemica più usata nelle discussioni accademiche contemporanee è quella elaborata dalla filosofa Miranda Fricker. Fricker (2007) teorizza che esistano diversi tipi di ingiustizie epistemiche, che hanno come comune denominatore un danno epistemico a carico di soggetti, siano essi singoli o gruppi, appartenenti a categorie sociali storicamente oppresse quali donne, persone di etnia non bianca o persone con malattie mentali. Il termine ingiustizia mette in rilievo come il danno non sia epistemicamente giustificato ma il frutto di pregiudizi inconsci contro l'identità razziale, sessuale, psicologica o

socioeconomica del soggetto che li subisce. I perpetratori, in genere inconsapevoli, di ingiustizia epistemica appartengono invece a gruppi sociali dominanti, quali uomini, eterosessuali, persone di etnia bianca o, in ambito psichiatrico, medici o ricercatori.

Analisi sistematiche di come l'IA applicata alla ricerca e alla cura psichiatriche possa impattare l'ingiustizia epistemica sono ancora allo stadio primordiale. In quel che segue, cercherò di cominciare a colmare questa lacuna. Innanzitutto, offrirò una breve panoramica dei tipi di IA esistenti (§2.1) e del loro funzionamento (§2.2). Spiegherò quindi che cosa si intende quando si parla di ingiustizia epistemica (§3), e in particolare di ingiustizia epistemica in psichiatria (§3.2), per poi restringere il focus al rapporto tra ingiustizia epistemica e IA (§4). Infine, cercherò di meglio delineare la relazione che intercorre tra ingiustizia epistemica e IA nella ricerca e nella cura della malattia mentale, delineando i rischi e le potenzialità offerti da questa tecnologia (§5).

2. Intelligenza Artificiale

2.1. *Tipi di Intelligenza Artificiale*

Esistono tre categorie di IA, che si distinguono in base alle loro abilità cognitive (IBM Data and AI team 2023):

- La *Narrow Artificial Intelligence*, anche nota come *Weak AI*, è una categoria di IA che possiede solo un tipo specifico di abilità cognitive. Previo addestramento da parte dell'uomo, essa è in grado di assolvere compiti che richiedono l'esercizio esclusivo di tali abilità. Al momento, è l'unico tipo di IA esistente.
- La *General Artificial Intelligence*, nota anche come *Strong AI*, è per il momento non più di un'idea. È un tipo di Intelligenza Artificiale paragonabile a quella umana, in grado di assolvere compiti intellettuali in contesti diversi senza la supervisione dell'uomo.
- La *Super Artificial Intelligence* è, al pari della General Artificial Intelligence, puramente teorica. È un tipo di IA le cui abilità cognitive sono superiori a quelle umane.

2.2. *Come funziona la (Weak) AI?*

Da ora in avanti, ogni volta che utilizzerò il termine IA farò riferimento alla Weak AI, che, come detto, è l'unico tipo di IA esistente al momento. In quel che segue, ne illustrerò alcune caratteristiche, pertinenti all'applicazione presente e futura di questi sistemi nella cura e nella ricerca sulla malattia mentale.

Per assolvere i compiti che le vengono assegnati, la IA sfrutta in genere tecniche di *machine learning* (ML). Lo scopo delle tecniche di ML è quello di addestrare programmi in grado di trovare delle leggi o funzioni che leghino determinati input a determinati output (UC Berkeley School of Information 2020). Uno dei modelli più diffusi di ML è il *deep learning* (DL). Il DL utilizza reti neurali artificiali la cui struttura rassomiglia a quella delle connessioni neuronali umane (Google Cloud, n.d.*).* Più specificamente, tali reti sono organizzate in livelli: i nodi di ciascun livello hanno il compito di apprendere una caratteristica specifica dei dati, in un crescendo di complessità e astrazione. Ad esempio, in una rete neurale di tre livelli per il riconoscimento delle immagini, il primo livello di nodi potrebbe essere dedicato all'identificazione dei bordi, il secondo a quella delle figure e il terzo a quella degli oggetti. Nel processo di addestramento della rete, affinché i dati possano essere meglio classificati e la rete possa performare al meglio, i pesi delle connessioni tra i nodi vengono modificati tramite una serie di tecniche quali *supervised*, *unsupervised* e *reinforcement* learning. Nel supervised learning, i programmatori dividono i dati in input e output e addestrano il programma a trovare le funzioni che legano i primi ai secondi, di modo che esso sia poi in grado di predire automaticamente output a partire da nuovi input. Nel reinforcement learning, in aggiunta alle fasi del supervised learning, il programma riceve un incentivo ogni volta che predice correttamente un output a partire da un nuovo input. Infine, nell'unsupervised learning i dati non sono previamente classificati in input e output: il programma tenta quindi di distinguere autonomamente i primi dai secondi e di trovare una funzione che li leghi, utilizzando criteri di prossimità e di raggruppamento quali il *clustering*.

Indipendentemente dai dettagli più tecnici, due aspetti essenziali del funzionamento dei metodi di ML sono la *natura statistica* di tali metodi

e la *qualità dei dati* con cui l'IA viene allenata. Tali caratteristiche saranno fondamentali per spiegare alcuni aspetti della relazione tra IA e ingiustizia epistemica — anche in ambito psichiatrico — che sarà l'oggetto dei prossimi capitoli.

3. Ingiustizia epistemica

3.1. *Tipi di ingiustizia epistemica*

Tra i tipi di ingiustizie epistemiche evidenziate da Fricker, di particolare rilevanza sono l'ingiustizia testimoniale e quella ermeneutica.

L'ingiustizia testimoniale è definibile come l'ingiustificato discredito del contributo o della testimonianza epistemica di una persona non in virtù della violazione di regole epistemiche da parte di quest'ultima, ma a causa di pregiudizi contro la sua identità. Tali sono i casi delle donne che a pari o superiori qualificazioni e competenze dei colleghi maschi, non vengono promosse, o delle persone che soffrono di malattie mentali, ritenute testimoni inaffidabili e irrazionali.

L'ingiustizia ermeneutica riguarda invece il silenziamento o la misinterpretazione del vissuto di categorie sociali non dominanti, le cui esperienze non rientrano nell'orizzonte concettuale ed emotivo dei gruppi sociali dominanti. Tra i danni primari delle ingiustizie ermeneutiche figurano il mancato riconoscimento delle esperienze di determinati gruppi sociali e la privazione, ai danni di tali gruppi, delle risorse intellettuali per poter meglio comprendere, elaborare ed esprimere queste esperienze. Fricker adduce a tal proposito il concetto di molestia sessuale, che non esisteva prima degli anni Settanta del Novecento, perché la testimonianza femminile veniva considerata irrilevante o distorta da categorie di giudizio prettamente maschili.

È interessante notare che le ingiustizie epistemiche sono spesso intersezionali, ovvero che arrecano un danno epistemico a soggetti appartenenti contemporaneamente a più di un gruppo sociale oppresso. Ad esempio, la testimonianza e le esperienze di una donna di colore affetta da una malattia mentale possono essere oggetto di ingiustizia epistemica a causa della sua identità sessuale, etnica e sociale contemporaneamente.

3.2. *Ingiustizia epistemica nella ricerca e nella cura della malattia mentale*

È noto come nell'immaginario comune le persone che soffrono di malattie mentali siano spesso ingiustamente associate a irrazionalità e inaffidabilità (Parcesepe e Cabassa 2013; Rössler 2016). Meno noto invece è che esse siano vittime di frequenti ingiustizie epistemiche per mano di chi meglio dovrebbe comprenderle e sostenerle, ovvero medici, psichiatri e ricercatori in campo psichiatrico. Secondo Crichton, Carel e Kidd (2017), uno dei primi studi specifici sul tema, le persone che soffrono di malattie mentali sarebbero particolarmente a rischio di subire un'ingiustizia epistemica in ambito psichiatrico a causa di tre fattori: 1) gli effetti dei disturbi mentali sulle capacità cognitive e interpersonali; 2) la preferenza data a concetti e linguaggi medici nel contesto dei discorsi sulla malattia e la salute mentale a discapito del linguaggio adottato dai pazienti; 3) stereotipi negativi sulle persone che soffrono di malattie mentali, diffusi anche in ambito medico. Le ingiustizie epistemiche a danno delle persone affette da disturbi mentali si declinerebbero nell'ambito medico sia nella variante testimoniale che in quella ermeneutica. Il contributo epistemico di tali persone sarebbe infatti ingiustamente screditato anche quando si tratta di meglio comprendere e curare la malattia mentale, rendendoli vittime di ingiustizia testimoniale. Similmente, poiché il linguaggio medico e scientifico viene epistemicamente privilegiato nella descrizione dei disturbi delle malattie mentali a discapito dell'esperienza soggettiva dei pazienti, questi sono molto spesso oggetto di ingiustizia ermeneutica. Il grado di ingiustizia epistemica subito dai pazienti psichiatrici varia in base a diversi fattori, quali la diagnosi, con un maggior grado di ingiustizia epistemica perpetuato ai danni di chi soffre di deliri cognitivi o di condizioni che hanno una forte componente cognitiva. Al capo opposto dello spettro, invece, si colloca chi è vittima di ingiusta depatologizzazione, che si verifica quando la severità di un disturbo viene sottovalutata e ridotta a caratteristiche o stranezze della personalità: il disturbo ossessivo–compulsivo è in questo senso emblematico.

Secondo Anastasia Scrutton (2017), una via per mitigare l'ingiustizia epistemica subita dai pazienti psichiatrici in ambito medico è quella di dare spazio alle loro esperienze e al loro vissuto di malattia mentale, a cui, è bene ricordare, essi hanno un accesso epistemico privilegiato

(*first-person authority*). Questo non significa accettare acriticamente qualsiasi contributo epistemico del paziente, ma renderlo parte attiva del processo di diagnosi e di cura insieme al medico. Il risultato di prendere sul serio le esperienze di chi vive la malattia mentale avrebbe ripercussioni positive non solo sul paziente, ma anche sulla ricerca e la cura psichiatrica, permettendo una comprensione più approfondita dei disturbi psichiatrici e una migliore personalizzazione del percorso di cura. Particolare attenzione va infine prestata al linguaggio del paziente, che, come accennato prima, è spesso screditato perché non conforme ai canoni del linguaggio medico–scientifico prevalente nella descrizione e concettualizzazione dei disturbi mentali, ma che è invece fondamentale per meglio capirli e curarli.

Altre proposte di soluzioni che sono state avanzate al problema dell'ingiustizia epistemica in ambito psichiatrico (Kidd, Spencer e Carel 2025) esortano a una definizione più precisa della nozione di *giustizia* epistemica, che non può essere ridotta alla mera assenza di ingiustizie; al coinvolgimento nel processo di ricerca e cura delle malattie mentali di tutti gli attori appartenenti alla sfera sociale del paziente, inclusi il paziente stesso, membri familiari, amici, assistenti sociali e attivisti; all'adozione di una visione strutturale delle soluzioni migliorative, supportate da una progettazione realistica; alla promozione di un cambiamento di paradigma culturale circa la malattia mentale, senza il quale cambiamenti più locali rimarrebbero vani.

4. Ingiustizia epistemica e IA

Qual è dunque il rapporto che intercorre tra ingiustizia epistemica e IA? Essendo l'IA uno strumento, la risposta più ovvia a questa domanda sembrerebbe essere che essa è neutrale in relazione all'ingiustizia epistemica, e che il suo maggiore o minor grado di ingiustizia dipenderebbe da come la si usa o la si programma.

A ben vedere, però, la situazione è più complicata di quel che potrebbe sembrare. Secondo Kay, Kasirzadeh e Mohamed (2024), dalla letteratura sul tema si possono delineare quattro diverse forme di ingiustizia epistemica specificamente collegate all'IA:

Amplificazione di ingiustizia testimoniale. I programmi di ML sono essenzialmente strumenti statistici, che predicono l'output più probabile a partire da determinati input. Naturalmente, più il campione di dati sarà ampio, maggiore sarà l'accuratezza e la rilevanza epistemica dell'output. Tuttavia, come giustamente notato da alcuni autori (Miragoli 2024; Bender *et al.* 2021), questo implica che i sistemi di IA basati sul ML (ovvero, una grandissima parte dei sistemi di IA) saranno soggetti a *conformismo epistemico*. Il conformismo epistemico consiste nel "trattare una data informazione come epistemicamente rilevante solamente perché essa è statisticamente dominante" (Miragoli 2024, p.10). Come si può facilmente intuire, il conformismo epistemico dell'IA favorisce la diffusione delle informazioni virtuali statisticamente dominanti, che in tal caso coincidono con quelle prodotte dai gruppi sociali dominanti, in genere maschi bianchi sufficientemente ricchi da avere accesso a internet. Basti pensare che più del 60% degli utenti di Reddit negli Stati Uniti sono uomini di un'età compresa tra i 18 e i 29 anni, e che solamente il 15% circa degli autori di Wikipedia sono donne. Non è necessariamente un male, dal punto di vista epistemico, che le informazioni statisticamente dominanti siano quelle che vengono diffuse dall'IA, poiché questo potrebbe portare, per esempio, all'esclusione di ideologie violente di carattere minoritario. Tuttavia, il principio statistico su cui si basa l'IA potrebbe anche intossicare l'ambiente epistemico, non solo favorendo la diffusione di disinformazione, qualora essa dovesse diventare statisticamente dominante online, ma anche promuovendo il patrimonio epistemico dei gruppi sociali dominanti a discapito di quello dei gruppi sociali minoritari, in un processo di *calcificazione epistemica* (Hardalupas 2024). Un nodo centrale della questione è inoltre che conformismo e calcificazione epistemiche sembrano essere caratteristiche non accidentali ma congenite dell'IA, che, operando secondo logiche statistiche, è strutturalmente disegnata per amplificare i patrimoni epistemici dominanti. Tale fenomeno è ulteriormente esacerbato dal fatto che alcune esperienze non sono facilmente quantificabili dal punto di vista algoritmico, e quindi a rischio di essere escluse dal patrimonio epistemico "artificiale".

1. *Manipolazione di ingiustizia testimoniale*. La propagazione di forme di ingiustizia testimoniale può essere non solo accidentale, perché già

presente nei dati e nella visione del mondo con cui la macchina viene allenata, ma anche intenzionale, promossa da attori malevoli per silenziare e opprimere il contributo epistemico di gruppi sociali non dominanti. È questo, ad esempio, il caso in cui attori malintenzionati sfruttano l'incerta distinzione tra testimonianze reali e testimonianze fabbricate dall'IA per screditare individui o intere comunità. In quest'ottica si spiegano l'intervento di un candidato americano al Congresso che proclamava la falsità dei video online documentanti la morte di George Floyd per mano della polizia, e la guida per creare meme suprematisti aggirando i filtri di sicurezza che circolava sulla piattaforma 4chan.

2. *Ignoranza ermeneutica di tipo generativo*. Diversamente dagli esseri umani, l'IA manca di conoscenza derivante dall'esperienza vissuta e di consapevolezza culturale. Questo porta inevitabilmente tali sistemi a dare una rappresentazione falsata del patrimonio esperienziale e concettuale delle minoranze. A differenza delle ingiustizie e dell'ignoranza ermeneutiche perpetrate dagli esseri umani, che possono essere corrette attraverso l'educazione e lo scambio di idee, quelle causate dall'IA sono congenite alla macchina stessa. È a causa di ignoranza ermeneutica congenita che un programma generatore di immagini come *Midjourney* può creare un'immagine di personaggi storici come nativi americani e guerrieri feudali giapponesi in posa per una foto che sorridono da un orecchio all'altro, un'attitudine chiaramente americana ma estranea alle popolazioni rappresentate.
3. *Ingiustizia ermeneutica di accessibilità*. I bias di tipo identitario insiti nell'IA possono anche ostacolare l'accesso di gruppi minoritari alle nuove tecnologie, generando forme di ingiustizia ermeneutica. Ad esempio, i programmi di riconoscimento vocale automatico faticano a riconoscere la parlata inglese di persone afroamericane, con il risultato che tali utenti devono adattare i loro pattern enunciativi o rinunciare all'uso di queste tecnologie. Similmente, poiché l'IA è sostanzialmente anglocentrica, i non anglofoni hanno difficoltà ad accedere ad informazioni di qualità nella loro lingua madre.

Altri rischi epistemici legati all'IA, che possono contribuire ad alimentare le suddette ingiustizie, sono *l'inondamento epistemico*, quando l'utente non è più in grado di distinguere le informazioni veritiere da quelle false a

causa dell'eccessiva quantità di informazioni; la *frammentazione epistemica*, quando le persone con esperienze e vissuti non dominanti non riescono a unirsi in una comunità epistemica a causa della marginalizzazione di questi vissuti da parte dell'IA; il *soffocamento testimoniale*, qualora un agente epistemico reprima il proprio contributo perché sente che l'ambiente epistemico in cui si trova non lo valorizzerebbe sufficientemente.

5. IA e ingiustizia epistemica nella ricerca e nella cura della malattia mentale: rischi e opportunità

È giunto il momento di affrontare la questione centrale di questa riflessione, ovvero come l'uso dell'IA in ambito psichiatrico potrebbe influenzare l'ingiustizia epistemica, nel bene e nel male.

5.1. *Rischi*

5.1.1. IA come massima autorità epistemica

Si è già trattato di come l'IA potrebbe essere indebitamente considerata la massima autorità epistemica nei processi decisionali riguardanti la cura dei pazienti, e di come questo rischio sia alimentato da una eccessiva deferenza di alcuni clinici nei confronti dell'IA. Questo atteggiamento potrebbe andare a detrimento del contributo epistemico sia dei clinici stessi che dei pazienti: in particolare, il contributo epistemico dei pazienti psichiatrici nel percorso di cura potrebbe rimanere ancor più inascoltato qualora si credesse che le informazioni derivanti dagli algoritmi rappresentino una forma di conoscenza superiore rispetto a tutte le altre. Questa eventualità potrebbe verificarsi, per esempio, qualora si privilegiassero le informazioni derivanti da fenotipizzazione digitale e *sentiment analysis* a discapito delle esperienze soggettive di malattia mentale.

5.1.2. Biased datasets

Similmente, forme di ingiustizia testimoniale si potrebbero verificare qualora i datasets utilizzati per allenare i programmi di IA fossero

inquinati da fattori quali etichette stigmatizzanti utilizzate per classificare dati collegati alla malattia mentale e una indebita correlazione della severità della malattia con l'etnia[1]. Anche in questi casi, l'apporto epistemico del paziente potrebbe essere sottovalutato a favore degli output generati dall'IA a partire da dati epistemicamente controversi.

5.1.3. Marginalizzazione di esperienze non quantificabili

Alcune esperienze di malattia mentale non facilmente quantificabili potrebbero essere escluse dal patrimonio epistemico artificiale, con una conseguente marginalizzazione del contributo epistemico di alcuni pazienti.

5.1.4. Marginalizzazione del patrimonio linguistico ed esperienziale dei pazienti

La diffusione dell'IA nella cura della malattia mentale potrebbe anche dare origine a ingiustizie di tipo ermeneutico, qualora il programma privilegiasse linguaggi e concettualizzazioni di tipo medico–scientifico a discapito del bagaglio linguistico e concettuale dei pazienti. Essendo l'IA strutturalmente improntata al conformismo epistemico, è molto probabile che questo rischio si concretizzi, dato che i linguaggi e la concettualizzazione medico–scientifica della malattia mentale sono nettamente prevalenti rispetto a quelli soggettivi dei pazienti. Per esempio, i sistemi di IA potrebbero generare ingiustizia ermeneutica tramite la classificazione algoritmica di diagnosi psichiatriche e previsioni di probabilità circa i benefici di una determinata cura, che potrebbero ulteriormente sopraffare le voci dei pazienti che non si conformano ai linguaggi e agli approcci terapeutici prevalenti in ambito medico.

(1) Storicamente, si registra un grado di severità di alcune malattie mentali più elevato tra le persone di etnia non bianca. Tuttavia, questo dato è principalmente dovuto al fatto che le minoranze incontrano più frequentemente delle barriere che ne riducono l'accesso ai servizi di salute mentale, e vengono intercettati quando la malattia è ormai a uno stadio avanzato.

5.1.5. Ingiustizia partecipativa

L'ingiustizia partecipativa è una forma di ingiustizia testimoniale in cui una persona viene considerata una fonte attendibile di informazioni, ma non di contributi epistemici che possano avanzare la conoscenza, quali opinioni, ipotesi o riflessioni critiche. Secondo Pozzi e De Proost (2024) una mancata partecipazione di persone che soffrono di malattie mentali al design e alla revisione di programmi di IA per la salute mentale quali chatbots, esporrebbe i malati mentali a un'ingiustizia partecipativa. Essi si ridurrebbero infatti a meri oggetti di studio e non sarebbero partecipanti attivi nel processo di sviluppo epistemico dei programmi.

Emblematico in questo senso sembra essere il caso, descritto dagli autori, del chatbot Karim, creato per offrire supporto psicologico ai profughi siriani, ma disegnato e implementato senza la partecipazione di alcun profugo. A questo dato si aggiunge la considerazione che Karim è l'adattamento di Tessa, chatbot disegnato per supportare persone che soffrono di ansia e depressione moderata negli Stati Uniti, quindi con un'esperienza e un background culturale molto diversi da quello dei profughi. I profughi siriani sarebbero dunque stati vittima di ingiustizia partecipativa, poiché trattati come oggetti piuttosto che come produttori di conoscenza.

5.2. *Opportunità*

L'IA potrebbe tuttavia anche rappresentare un'opportunità per contrastare con maggiore efficacia l'ingiustizia epistemica già presente in ambito psichiatrico.

5.2.1. Individuazione di ingiustizia testimoniale nei dati virtuali

Con l'aiuto delle persone che soffrono di malattie mentali, l'IA può essere usata per individuare su larga scala i contesti, i perpetratori e la prevalenza dell'ingiustizia testimoniale online. In particolare, essa potrebbe servire a rilevare quelle narratizioni che chi soffre di malattia mentale considera epistemicamente ingiuste in un *corpus* più o meno ampio di dati.

5.2.2. Generazione di risorse ermeneutiche

L'IA può anche essere usata per l'espressione di esperienze lontane da quelle comuni. Ad esempio, un paziente affetto da schizofrenia potrebbe trovare utile comunicare le proprie allucinazioni ai medici o ai familiari attraverso un programma di generazione di immagini. Similmente, come recentemente notato da Elisabetta Lalumera (2024), la fenotipizzazione digitale (i.e., l'identificazione di pattern comportamentali osservabili quali risposte emotive, processi cognitivi, movimenti, a partire da dati virtuali, quali frequenza cardiaca, risposte a questionari sul benessere psicologico, raccolti tramite l'interazione passiva o attiva del paziente con strumenti tecnologici), un particolare tipo di IA, potrebbe servire a combattere ingiustizie testimoniali ed ermeneutiche in ambito psichiatrico. L'esempio fittizio creato da Lalumera ha al centro A, una professoressa universitaria di quarant'anni, divorziata con due figli, che per tutta la sua vita ha sofferto di disturbi di attenzione, episodi depressivi, fatica nel regolare le proprie emozioni e impulsività nella gestione dei soldi. Leggendo, A viene a conoscenza dell'esistenza di ADHD (Attention Deficit Hyperactivity Disorder), un disturbo psichiatrico nei cui sintomi riconosce molti dei suoi meccanismi emotivi e cognitivi. A prende quindi appuntamento da uno psicologo, ma quest'ultimo dismette l'ipotesi della donna perché ella ha una vita professionale e privata soddisfacente, e a prima vista non sembra mostrare segni di sofferenza psichica. Passano gli anni, e nascono degli strumenti di fenotipizzazione digitale in grado di rilevare i sintomi di ADHD negli adulti. A si sottopone ai nuovi test e risulta positiva a ADHD, confermando i sospetti che aveva avuto in passato. In questo caso, secondo Lalumera, la fenotipizzazione digitale contribuisce a mitigare sia l'ingiustizia testimoniale di cui A era stata vittima per mano dello psicologo (che aveva dismesso il contributo epistemico di A in base allo stereotipo che i pazienti non posseggono autorità epistemica sulle loro malattie e a quello che una persona malata non può avere successo nella vita professionale e privata), sia l'ingiustizia ermeneutica, validando la narrativa e i sentimenti di A sulla propria malattia. Lalumera risponde anche all'obiezione che la fenotipizzazione digitale sia sostanzialmente neutrale nei confronti dell'ingiustizia epistemica, mitigandola quando conferma il

racconto del paziente e amplificandola quando lo smentisce. Secondo Lalumera, l'IA può diventare uno strumento di oppressione epistemica solamente qualora a essa sia attribuita una priorità epistemica non giustificata, ovvero indipendentemente dal suo livello di accuratezza. Se a un test di gravidanza, la cui accuratezza è del 99%, è razionale attribuire priorità epistemica rispetto alle sensazioni del paziente, non è invece ragionevole ascriverla agli attuali strumenti di IA impiegati in ambito psichiatrico, la cui accuratezza è molto difficile da stimare. In secondo luogo, anche qualora l'IA per la ricerca e la cura della malattia mentale raggiungesse una accuratezza maggiore, essa non potrebbe comunque essere considerata l'ultima autorità epistemica nella scelta del piano terapeutico, che deve sempre nascere da un dialogo tra medico e paziente, e tenere in conto dei desideri, valori e necessità di quest'ultimo. In conclusione, secondo Lalumera, l'ingiustizia epistemica non sarebbe connaturata alla fenotipizzazione digitale, ma dipenderebbe dalla mancata comprensione del funzionamento di queste nuove tecnologie e dei loro output.

6. Conclusione

L'applicazione dell'IA alla ricerca e alla cura della malattia mentale è ancora agli esordi, ma emergono già alcune caratteristiche di questi sistemi che potrebbero esacerbare preesistenti ingiustizie epistemiche di tipo testimoniale, ermeneutico e partecipativo. In particolare, l'IA sembra presentarsi come una forza conservatrice e amplificatrice del patrimonio epistemico dei gruppi sociali dominanti, a discapito dei patrimoni epistemici dei gruppi sociali svantaggiati. Per quanto riguarda la ricerca e la cura della malattia mentale, questo si traduce in un'amplificazione del patrimonio epistemico medico–clinico a discapito di quello esperienziale dei pazienti, con importanti ripercussioni di tipo morale ed epistemico, culminanti in forme di ingiustizia testimoniale, ermeneutica e partecipativa.

Parallelamente, l'IA potrebbe anche rivelarsi un utile strumento per combattere l'ingiustizia testimoniale ed ermeneutica che già affligge chi soffre di malattie mentali, generando nuove risorse ermeneutiche

e contribuendo a individuare narrazioni epistemicamente distorte sulla malattia mentale nei dati online.

La via più promettente per limitare i rischi e massimizzare le potenzialità epistemiche dei sistemi di IA in ambito psichiatrico sembra essere quella di suscitare nei clinici che li utilizzano e nei tecnici che li programmano una maggiore consapevolezza dei meccanismi e dei pericoli dell'ingiustizia epistemica legati alle nuove tecnologie, insieme a un serio coinvolgimento dei malati nella fase di design di tali strumenti.

Riferimenti bibliografici

Allen B., S. Agarwal, L. Coombs, C. Wald e K. Dreyer (2021) *2020 ACR Data Science Institute Artificial Intelligence Survey*, "Journal of American College of Radiologists", 8(11): 53–59.

Bender E.M., T. Gebru, A. McMillan–Major, e S. Shmitchell (2021) *On the Dangers of Stochastic Parrots. Can Language Models Be Too Big?*, "Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery", 610–623.

Berry D.M. (2018) "Weizenbaum, ELIZA and the End of Human Reason", in M. Baranovska e S. Höltgen (a cura di), *Hello, I'm Eliza: Fünfzig Jahre Gespräche mit Computern*, Projekt Verlag, Berlin, 53–70.

Cheng M. (2024) *FDA Clears First Digital Treatment for Depression, but Experts Caution That Research Is Still Early*, CNN. https://edition.cnn.com/2024/04/02/health/fda-rejoyn-depression-digital-treatment/index.html (ultimo accesso, 26 febbraio 2025).

Crichton P., H. Carel e I.J. Kidd (2017) *Epistemic Injustice in Psychiatry*, "BJ Psychiatry Bulletin", 41(2): 65–70.

Fleuren L.M., P. Thoral, D. Shillan, A. Ercole e A. Elbers (2020) *PWG, Right Data Right Now Collaborators. Machine Learning in Intensive Care Medicine: Ready for Take-off?*, "Intensive Care Medicine", 46:1486–1488.

Frana P.L. e M.J. Klein (2021) *Encyclopedia of Artificial Intelligence: The Past, Present, and Future of AI*, Bloomsbury, London.

Fricker M. (2007) *Epistemic Injustice: Power and the Ethics of Knowing*, Oxford University Press, Oxford.

Google Cloud. *Che cos'è il deep learning?* https://cloud.google.com/discover/what-is-deep-learning (ultimo accesso 25 febbraio 2025).

Hardalupas M. (2024) *Contributory Injustice, Epistemic Calcification, and the Use of AI Systems in Healthcare*, "Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024)", 573–583.

IBM Data e AI team (2023) *Understanding the Different Types of Artificial Intelligence*, IBM, https://www.ibm.com/think/topics/artificial-intelligence-types. (ultimo accesso 25 febbraio 2025)

Kay J., A. Kasirzadeh e S. Mohamed (2024) *Epistemic Injustice in Generative AI*, "Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES '24)", 684–697.

Khaku Z. (2023) *Limbic*, NHS. https://nhsaccelerator.com/innovation/limbic/ (ultimo accesso 26 febbraio 2025)

Kidd I.J., L. Spencer e H. Carel (2025) *Epistemic Injustice In Psychiatric Research And Practice*, "Philosophical Psychology", 38(2): 503–531.

Lalumera E. (2024) "Ameliorating Epistemic Injustice with Digital Health Technologies", in L. Bortolotti (a cura di) *Epistemic Justice in Mental Healthcare*, Palgrave Macmillan, London, 141–158.

Mankins J.C. (1995) *Technology Readiness Levels. A White Paper*, NASA, Washington (DC).

McAllen J. (2024) *The Therapist in the Machine*, "The Baffler", 76, https://thebaffler.com/salvos/the-therapist-in-the-machine-mcallen.

Miragoli M. (2025) *Conformism, Ignorance & Injustice: AI as a Tool of Epistemic Oppression*, "Episteme", 22(2): 522-540. doi:10.1017/epi.2024.11.

Monteith S., T. Glenn, J. Geddes, P.C. Whybrow, E. Achtyes e M. Bauer (2022) *Expectations for Artificial Intelligence (AI) in Psychiatry*, "Current Psychiatry Report", 24: 709–721.

Parcesepe A.M. e L.J. Cabassa (2013) *Public Stigma of Mental Illness in the United States: A Systematic Literature Review*, "Administrative Policy for Mental Health", 40(5): 384–399.

Pozzi G. e M. De Proost (2024) *Keeping an AI On the Mental Health of Vulnerable Populations: Reflections on the Potential for Participatory Injustice*, "AI Ethics", 5: 2281–2291.

Rössler W. (2016) *The Stigma of Mental Disorders: A Millennia–Long History of Social Exclusion and Prejudices*, "EMBO Reports", 17(9): 250–253.

Scrutton A.P. (2017) "Epistemic Injustice and Mental Illness", in I.J.

Kidd, J. Medina e G. Pohlhaus Jr. (a cura di), *The Routledge Handbook to Epistemic Injustice*, Routledge, London, 347–355.

Solow R. (1987, July 12th) *We'd Better Watch Out*, "New York Times", 36, https://gwern.net/doc/economics/automation/1987-solow.pdf (ultimo accesso 25 febbraio 2025).

UC Berkeley School of Information (2020) *What Is Machine Learning (ML)?*, https://ischoolonline.berkeley.edu/blog/what-is-machine-learning/#:~:text=Definition%20of%20Machine%20Learning,machine%20learning%20via%20R2D3%20open_in_new) (ultimo accesso 25 febbraio 2025).

Van de Sande D., M.E. van Genderen, J. Huiskens, D. Gommers e J. van Bommel (2021) *Moving from Bytes to Bedside: A Systematic Review On The Use Of Artificial Intelligence In The Intensive Care Unit*, "Intensive Care Medicine", 47:750–760.

Weizenbaum J. (1966) *ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine*, "Communications of the ACM", 9(1): 36–45.