Empatia artificiale e dialettiche della vulnerabilità. Ripensare la psicoterapia nell'era digitale

ANTONIO CARNEVALE, ALESSIA CORREANI E PAOLO QUAGLIARELLA [*]

DOI: https://doi.org/10.15162/1827-5133/2237

ABSTRACT

L'articolo analizza in modo critico l'introduzione dei *Large Language Models* (LLM) nell'ambito della psicoterapia e della salute mentale, esplorando le implicazioni filosofiche, neuroscientifiche ed etiche della crescente integrazione umano-macchina. L'indagine si concentra sulla dialettica tra corpo, mente e tecnologia, mettendo in discussione il dualismo tradizionale e sottolineando come gli avanzamenti tecnologici stiano riconfigurando profondamente il rapporto tra vulnerabilità, cura e potenziamento umano. Sebbene i chatbot terapeutici offrano nuove opportunità di accesso immediato e continuo al supporto psicologico, gli autori mettono in guardia rispetto a rischi quali la simulazione di empatia, la cristallizzazione del disagio, l'autoterapia solipsistica e i *bias* algoritmici intrinseci. L'articolo propone un approccio terapeutico ibrido e una governance etica rigorosa come strumenti necessari per valorizzare le potenzialità degli LLM senza perdere di vista la centralità irrinunciabile della relazione umana nella psicoterapia. In conclusione, si auspica una riflessione critica e responsabile sul ruolo dell'IA, sottolineando la necessità di un quadro etico-operativo capace di prevenire derive riduzionistiche e sfruttamenti commerciali della vulnerabilità umana.

[*] Antonio Carnevale è ricercatore presso il dipartimento DIRIUM dell'Università degli Studi di Bari Aldo Moro e co-fondatore di DEXAI srl.

Alessia Correani è direttrice di Digital @Sky Italia, Ethics Advisor e docente di User Experience Psychology presso l'Università Cattolica di Milano.

Paolo Quagliarella è filosofo, psicologo a orientamento analitico archetipico e docente di Epistemologia II presso la Scuola di Psicoterapia analitico archetipica Atanor, IT Business Analyst in Pharma Quality Europe.

La redazione dei paragrafi è così distribuita: il primo paragrafo è stato scritto da Alessia Correani, il secondo da Paolo Quagliarella, il terzo da Antonio Carnevale.

The article analyses critically the introduction of Large Language Models (LLM) in the field of psychotherapy and mental health, exploring the philosophical, neuroscientific and ethical implications of the increasing human-machine integration. The investigation focuses on the dialectic between body, mind and technology, questioning traditional dualism and highlighting how technological advances are profoundly reconfiguring the relationship between vulnerability, care and human empowerment. Although therapeutic chatbots offer new opportunities for immediate and continuous access to psychological support, the authors warn about risks such as simulated empathy, crystallization of distress, solipsistic self-therapy and inherent algorithmic bias. The article proposes a hybrid therapeutic approach and a rigorous ethical governance as necessary instruments to enhance the potential of LLM not forgetting the inalienable centrality of the human relationship in psychotherapy. In conclusion, a critical and responsible reflection on the role of AI is called for, emphasising the necessity of an ethical-operational framework capable of preventing reductionist drifts and commercial exploitation of human vulnerability.

Mettiamo in discussione il classico dualismo cartesiano, che separa mente e corpo e identifica l'essere umano esclusivamente con il pensiero: "Cogito, ergo sum". Questo paradigma è oggi messo in dubbio sia dalla filosofia contemporanea, sia dalle più recenti scoperte scientifiche[1]. Il corpo umano, nella sua dimensione biologica e fisica, si è rivelato parte integrante ed essenziale dei processi psico-cognitivi di formazione dell'intelligenza, come dimostrano diversi filoni della ricerca neuroscientifica.

A esemplificare significativamente questa tendenza è la scoperta di un potente asse di interazione tra corpo e cervello che coinvolge e connette i sistemi nervosi centrale e periferico con il sistema nervoso enterico (o sistema nervoso dell'apparato gastroenterico)[2]. Un tipo di connessione che è noto come "gut-brain axis", ovvero "asse intestino-cervello"[3]. Il ruolo di tale asse è stato evidenziato tanto nello sviluppo cognitivo, quanto nell'influenza esercitata sul comportamento umano, fino ad essere riconosciuto come possibile attivatore o cofattore in alcuni disturbi psichici[4]. Stiamo parlando di un apparato intelligente che si estende anche ad altri organi vitali, come il cuore e l'asse ipotalamico-pituitario e adrenalinico che regola il livello di reattività del corpo in base a stimoli stressogeni esterni (paura, rischio). Si tratta perciò di un complesso sistema di comunicazioni neuronali distribuito che guiderebbe molti comportamenti cosiddetti "intelligenti" ed adattivi.

A sostegno di queste visioni connessioniste, studi recenti hanno individuato neuroni meccanocettori che innervano in modo sensitivo il cuore e lo stomaco, influenzando sia stati metabolici corporei sia dimensioni emotive come ansia e

[1] A titolo esemplificativo, si veda: A. Damasio, *L'errore di Cartesio: Emozione, ragione e cervello umano*, Adelphi, Milano 1994.

[2] Cfr. J. A. Foster e K-A. McVey Neufeld, "Gut–Brain Axis: How the Microbiome Influences Anxiety and Depression", in «Trends in Neurosciences», vol. XXXVI, 2013, n. 5, pp. 305-312, consultabile qui:
<https://doi.org/10.1016/j.tins.2013.01.005> (consultato il 24/04/25).

[3] E. Alim *et al*., "Enteric Nervous System, Gut-Brain Connection and Related Neurodevelopmental Disorders", in «Anatomy», vol. XIV, 2020, n. 1, pp. 61-67, consultabile qui:
<doi.org/10.2399/ana.20.008> (consultato il 24/04/25).

[4] Vedi E. Alim *et al*., op. cit.

depressione[5]. Tutto ciò rafforza una concezione integrata del rapporto tra cervello e corpo, in cui il sistema cognitivo-comportamentale umano appare profondamente influenzato dalla nostra struttura biologica e dalle interazioni dinamiche con l'ambiente fisico esterno. Un tale intreccio di livelli – biologico, neurologico, ambientale – richiama, per complessità e interdipendenza, il tipo di paradigma sistemico e relazionale che è al cuore della Teoria della Relatività Generale di Albert Einstein, dove nessun elemento può essere compreso in isolamento, ma solo in relazione al contesto in cui è immerso.

Stante questa conformazione dinamica della nostra integrità, è facile intuire perché l'emergere delle nuove tecnologie *digitali* – e in particolare, negli ultimi decenni, dell'intelligenza artificiale (IA) – possa tanto preoccupare e far discutere, cioè proprio a causa della straordinaria potenzialità che queste hanno di influenzare, potenziare, modificare o interferire con il comportamento e la cognizione umana. Soprattutto se pensiamo alle più recenti applicazioni dell'IA, come appunto i sistemi di *Large Language Models* (LLM) oggetto di questo articolo, ci troviamo di fronte alla manifestazione di capacità sia di apprendimento sia di "scorciatoia" che qualcuno non ha avuto paura di definire *aliena*[6]. La sua eccezionalità – questo il nostro parere – è tale poiché opera come intelligenza sia nel senso di controllo cognitivo-funzionale "verticale" sia come dominio relazionale "orizzontale". Per intelligenza artificiale "verticale" si intende quell'insieme di tecniche e tecnologie che utilizzano molti dati, li modellano e li operazionalizzano cercando di rispondere a esigenze molto specifiche di processamento di un dominio *funzionale*, ad esempio il processamento di stimoli visivi, la loro catalogazione e riconoscimento. Proprio come avviene nel sistema visivo umano, dove lo stimolo visivo viene trasdotto

[5] Cfr. "Mechanosensitive Neurons Innervating the Gut and Heart Control Metabolic and Emotional State", in «Nature Metabolism», 7, 2025, pp. 249-250, consultabile qui: <https://doi.org/10.1038/s42255-024-01208-3.> (consultato il 24/04/25).

[6] Nello Cristianini definisce l'intelligenza artificiale come una "scorciatoia" per ottenere risultati intelligenti senza replicare i meccanismi dell'intelligenza umana. In questa accezione, l'IA non mira a comprendere o imitare la coscienza, la razionalità o la comprensione, qualità che sono proprie degli esseri umani, ma sfrutta metodi alternativi – come l'elaborazione statistica di grandi quantità di dati – per risolvere problemi in modo efficace. La scorciatoia è dunque tecnica, non cognitiva: un modo per arrivare al comportamento "intelligente" senza necessariamente passare per i processi mentali che lo caratterizzano negli esseri umani. Cfr. N. Cristianini, *La Scorciatoia*, Il Mulino, Bologna 2023.

dalla retina e trasmesso attraverso il nervo ottico fino alle aree corticali deputate all'elaborazione e al riconoscimento dell'informazione visiva, l'intelligenza artificiale "verticale" opera attraverso una sequenza funzionale specializzata che elabora dati grezzi trasformandoli in rappresentazioni utilizzabili. Sul versante opposto, l'intelligenza artificiale "orizzontale" può essere definita come quella modalità di interazione intelligente che, diversamente da una specializzazione verticale focalizzata su compiti specifici, agisce attraverso domini ampi e diversificati, stabilendo connessioni trasversali fra aree differenti del sapere e del comportamento. Una simile forma di intelligenza, incarnata efficacemente dai sistemi di LLM, opera tramite una capacità adattiva di dialogo, interpretazione e produzione di significati che superano il semplice compito funzionale.

Questa peculiare abilità dei sistemi di IA nel trattare con la complessità umana, tanto in verticale quanto in orizzontale, agendo come piattaforme che facilitano una comunicazione fluida e versatile, capaci di integrare, sintetizzare e reinterpretare informazioni provenienti da contesti molto eterogenei (tanto da renderci tutte e tutti *simbionti* di interazioni umano-macchina sempre più avviluppanti[7]), trasforma questi strumenti in forze capaci di ridefinire radicalmente il nostro rapporto con la conoscenza e la comunicazione, ma non senza introdurre nuove vulnerabilità e dilemmi etici. Se, pensando al futuro, è infatti ipotizzabile che l'IA possa trasformarsi un giorno in un agente collaborativo capace di ampliare e potenziare le capacità cognitive umane, creando reti relazionali che favoriscano l'emergere di nuove idee e soluzioni creative, tuttavia è necessario rivolgere la riflessione all'oggi e capire in quali modi la tecnologia digitale e l'interazione dell'intelligenza umana con quella artificiale sia potenziante o de-potenziante e come in casi di inesperienza, ignoranza o vulnerabilità occorra intervenire e nei migliori casi prevenire.

Oggi numerose funzioni cognitive primarie dell'essere umano sono state emulate con successo dai sistemi di intelligenza artificiale. Già nel 2018, diversi studi empirici hanno attestato il raggiungimento di una parità prestazionale tra esseri umani e macchine in vari ambiti funzionali, quali la percezione visiva e il riconoscimento di immagini, l'elaborazione acustica e la trascrizione

[7] Cfr. A. Carnevale, *et. al.*, "A Human-Centred Approach to Symbiotic AI: Questioning the Ethical and Conceptual Foundation", in «Intelligenza Artificiale», 2024, pp. 1-12, consultabile qui: <https://doi.org/10.3233/IA-240034> (consultato il 24/04/25).

del parlato, la generazione linguistica e la scrittura[8]. Anche per quanto concerne le funzioni cognitive superiori, le tecnologie contemporanee sembrano in grado di simulare comportamenti umani con un grado crescente di accuratezza. Dalla traduzione automatica alla risoluzione di problemi multifattoriali, dalla generazione di contenuti linguistici alla formulazione di risposte complesse, dalla previsione di eventi all'identificazione di correlazioni significative entro grandi moli di dati: tutte queste operazioni, pur essendo il prodotto di processi computazionali formalizzati, mostrano livelli di performance comparabili – talvolta superiori – a quelli riscontrabili nella cognizione umana.

Certo, è necessario esercitare cautela nell'impiego del termine "comprensione" quando si fa riferimento alle prestazioni delle macchine. Utilizzare questa nozione senza ulteriori qualificazioni rischia infatti di attribuire a dispositivi computazionali caratteristiche che, tradizionalmente, rimandano a processi fenomenologici, intenzionali e incarnati inerenti alla soggettività umana[9]. Tuttavia, proprio in quanto atto fenomenologico, la comprensione non si esaurisce in un evento interno alla coscienza, ma si costituisce anche

[8] Tra gli studi più noti, ricordiamo quello sul riconoscimento vocale condotto da Microsoft. La società americana nel 2017 ha annunciato di aver raggiunto una precisione pari a quella umana nel trascrivere conversazioni telefoniche nel corpus *Switchboard*. Tuttavia, è importante notare che questi risultati si riferiscono a contesti specifici e controllati. Per un ulteriore approfondimento, rimandiamo a Y. Shoham, *et al.*, *The AI Index 2018 Annual Report*, AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford (CA) 2018, consultabile qui: <https://hai.stanford.edu/ai-index/2018-ai-index-report> (consultato il 24/04/2025). Si veda inoltre I. Beaver, "*Is AI at Human Parity Yet? A Case Study on Speech Recognition*", in «AI Magazine», vol. LXIII, 2022, n. 4, 2022, pp. 386-389.

[9] L'uso del termine "comprensione" in relazione alle macchine è al centro di un dibattito filosofico di lunga data, riacceso in modo emblematico dall'esperimento mentale della "stanza cinese" proposto nel 1980 da John Searle, volto a mostrare che un software può manipolare simboli secondo regole sintattiche senza per questo comprendere il significato di ciò che elabora. La discussione ha coinvolto numerosi filosofi della mente e teorici dell'intelligenza artificiale, contrapponendo posizioni come quella del cognitivismo computazionale – che interpreta la mente come un sistema di elaborazione simbolica – a visioni che sottolineano la distinzione tra sintassi e semantica. Tra le risposte più note vi è quella di D.C. Dennett, secondo cui l'intelligenza artificiale manifesta forme di "competenza" che, pur prive di consapevolezza, sono funzionalmente indistinguibili dalla comprensione umana. Si vedano J. Searle, "Minds, Brains and Programs", in «The Behavioral and Brain Sciences», vol. III, 1980, n. 3, pp. 417-457; D.C. Dennett, *The Intentional Stance*, MIT Press, Cambridge (MA) 1987; T. Horgan e J. Tienson, *Connectionism and the Philosophy of Psychology,* MIT Press, Cambridge (MA) 1996.

attraverso le pratiche, gli strumenti e gli ambienti che rendono possibile l'emergere di significati condivisi. In questo senso, gli artefatti tecnologici non si limitano a imitare la comprensione: essi contribuiscono a modellarne le condizioni di possibilità, intervenendo nella mediazione tra soggetto e mondo[10]. È perciò lecito chiedersi: se l'intelligenza artificiale incide già oggi, e con ogni probabilità inciderà ancor più radicalmente in futuro, sul nesso mente-corpo, contribuendo a ridefinire il rapporto tra individuo, tecnologia e società, quali sono le responsabilità a cui siamo chiamati? Quali forme di attenzione critica, quali dispositivi normativi e quali pratiche etiche saranno necessari per orientare l'evoluzione di sistemi capaci non solo di operare, ma di trasformare le condizioni stesse della nostra capacità di comprendere?

Se, per un verso, le tecnologie di IA forniscono un indubitabile vantaggio in quei settori commerciali che meglio sfruttano, più di altri, la scienza dei dati e i suoi sviluppi ingegneristici – quali manifatturiero, farmaceutico, scientifico, bellico, *retail* – tuttavia, per il verso opposto, tale vantaggio rappresenta uno svantaggio per chi, ad esempio, non ha una cultura aziendale adeguata o il denaro sufficiente per adeguarsi. Le aziende ed i paesi che hanno investito in acquisizione e gestione di importanti volumi di dati hanno goduto delle condizioni favorevoli per implementare i sistemi di *machine learning*, riuscendo così a snellire e semplificare processi aziendali molto costosi e a raggiungere vantaggi competitivi importanti come il monopolio di conoscenza e/o tecnologia. Per queste limitate realtà, è cresciuto il valore strategico di competitività, rendendosi più veloce il *go-to-market* e, al contempo, si sono anche ridotti i costi operativi e di personale, restituendo tempo al comparto di risorse umane nel *middle-management* impegnato a prendere decisioni operative importanti.

Ma in gioco non vi è soltanto una questione di responsabilità etica legata all'economia e all'innovazione tecnologica. Più radicalmente, è in discussione

[10] Sul potere di *agency* della tecnologia in forma di mediazione, cfr. P.-P. Verbeek, *What Things Do. Philosophical Reflections on Technology, Agency, and Design*, Penn State University Press, University Park (PA) 2005. Questa impostazione si inserisce in una più ampia corrente di studi post-fenomenologici nell'ambito della filosofia della tecnologia, che esplora il ruolo di mediazione degli artefatti tecnologici nei processi di costituzione dell'esperienza e della soggettività. Si veda in particolare R. Rosenberger e P.-P. Verbeek (a cura di), *Postphenomenological Investigations. Essays on Human-Technology Relations*, Lexington Books, London 2015.

il rapporto tra tecnologia e sapere, ovvero il tipo di relazione che viene a instaurarsi tra i dispositivi tecnici e le forme del conoscere, dell'apprendere, del ricordare. In questa prospettiva, ciò che rischia di venir meno è la distinzione critica tra un sé cognitivo, capace di interrogarsi sul proprio agire, e un sé tecnologico, sempre più immerso in automatismi operativi. La dialettica tra questi due poli – tra l'umano che conosce e l'umano che esternalizza tale conoscenza nella tecnica – si interrompe proprio quando smettiamo di chiederci come, e in base a quali categorie, la tecnologia contribuisca alla formazione di ciò che chiamiamo "conoscenza"[11].

Eppure, per quanto sofisticata, la tecnologia resta uno strumento concepito, progettato e implementato da esseri umani. Questo implica che le sue logiche interne, i suoi modelli decisionali e persino i suoi automatismi riflettano, in misura più o meno consapevole, le inclinazioni e i limiti cognitivi di chi la sviluppa. L'essere umano, infatti, tende per sua natura a semplificare, a risparmiare risorse mentali, e questo comporta scorciatoie cognitive che possono compromettere coerenza e precisione nei processi decisionali. Tali dinamiche si riflettono inevitabilmente anche nei sistemi che costruiamo. Se l'errore

[11] Su queste tematiche si innesta una lunga e articolata riflessione filosofica che, da prospettive differenti, ha interrogato il rapporto tra tecnica e sapere, sottolineando come ogni dispositivo tecnologico non sia mai neutro, ma incorpori visioni del mondo, gerarchie di valore e strutture di potere. Nella tradizione fenomenologica, M. Heidegger, *La questione della tecnica* (1953), goWare, Firenze 2017 ha evidenziato come la tecnica moderna, più che un semplice insieme di strumenti, rappresenti un modo specifico di disvelamento del reale, un orizzonte ontologico che predispone l'essere umano a considerare il mondo come pura riserva di risorse disponibili. In una direzione più costruttivista e processuale, G. Simondon, *Du mode d'existence des objets techniques*, Aubier-Montaigne, Paris 1958 ha posto l'accento sul carattere relazionale e generativo della tecnica, concependo l'oggetto tecnico come un elemento in divenire che partecipa attivamente alla costituzione del soggetto e dell'ambiente. Questa linea sarà poi ripresa e ampliata da B. Stiegler, *La technique et le temp, vol. I. La faute d'Épiméthée*, Galilée, Paris 1994, che interpreta la tecnica come memoria esterna e come condizione trascendentale della soggettività umana. In ambito femminista, D. Haraway, "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective", in «Feminist Studies», vol. XIV, 1988, n. 3, pp. 575-599 ha criticato la pretesa di oggettività neutrale propria della scienza moderna, proponendo una teoria della conoscenza "situata" che riconosce come i saperi siano sempre localizzati, incarnati e attraversati da relazioni di potere, incluse quelle di genere, razza e classe. Queste prospettive, pur nelle loro differenze, convergono nel mettere in discussione l'idea di una conoscenza disincarnata e universalizzabile, mostrando come il gesto tecnico sia sempre anche un gesto carico di essenzialità da svelare, intriso di senso epistemico, etico e politico.

umano è una componente insopprimibile della nostra intelligenza – e talvolta persino la condizione della scoperta – è allora essenziale sottoporre costantemente tali processi a revisione critica[12]. La nostra capacità di giudizio è infatti spesso minata da *bias* e pregiudizi radicati, che interferiscono con l'esercizio di un pensiero lucido e razionale. E poiché queste stesse distorsioni possono essere replicate e amplificate dagli algoritmi, occorre interrogarsi con urgenza sulle responsabilità morali, cognitive e sociali che gravano su chi sviluppa e impiega queste tecnologie.

Quali sono, allora, le responsabilità di coloro che progettano algoritmi di raccomandazione e predizione capaci di influenzare, in modo anche impercettibile, il comportamento individuale e collettivo? Su quali criteri decidono cosa sia rilevante, desiderabile, utile per l'utente? E ancora: coloro che occupano oggi le posizioni chiave nelle Big Tech – per lo più uomini, bianchi, di origine occidentale – possiedono davvero gli strumenti cognitivi, culturali ed etici per comprendere le conseguenze del proprio operato sullo sviluppo umano? La pervasività quotidiana degli algoritmi, l'interazione continua e spesso compulsiva con sistemi che orientano le nostre scelte, suggeriscono pensieri, offrono scorciatoie e stimolano desideri, interroga profondamente il nostro benessere psico-fisico. Che effetto ha tutto questo sulle nostre capacità attentive, mnestiche, linguistiche? Sulla nostra facoltà di sintesi, di profondità di pensiero, sulla *lentezza necessaria*[13] all'elaborazione critica e alla creatività? Se le macchine assorbiranno sempre più il "lavoro cognitivo pesante", che ne sarà di quel pensiero che richiede concentrazione, sforzo, immersione e passione? Di quel sapere che non si riduce all'efficienza ma scaturisce, talvolta, dalla fatica del pensare? [14]

[12] Cfr. T. Numerico, *Big data e algoritmi. Prospettive critiche*, Carocci, Roma 2021.
[13] Cfr. D. Kahneman, *Pensieri lenti e veloci*, Mondadori, Milano 2012.
[14] Cfr. M. Csíkszentmihályi, *Flow: The Psychology of Pptimal Experience*, Harper & Row, New York 1990.

Tecnologie affettive e chatbot terapeutici: il futuro emotivo della salute mentale

Nel panorama della salute mentale, l'utilizzo di tecnologie digitali e di strumenti di intelligenza artificiale sta assumendo un rilievo sempre più evidente. Un esempio emblematico di questa trasformazione è rappresentato dai modelli di LLM, spesso implementati in forma di chatbot, che sono in grado di interagire con gli utenti tramite un linguaggio naturale realistico. Diversi studi, condotti sia in ambito accademico sia nel settore privato, stanno esplorando le potenzialità di questi strumenti come supporto alla psicoterapia e alla promozione del benessere mentale[15].

Ma cosa sono gli LLM? Si tratta di sistemi di intelligenza artificiale addestrati su enormi quantità di dati testuali, spesso dell'ordine di miliardi di parole, con l'obiettivo di "imparare" le regole e le strutture del linguaggio umano[16]. Grazie a tecniche di *deep learning*, questi modelli sono in grado di generare risposte coerenti e contestualmente pertinenti. Quando tali modelli vengono impiegati all'interno di chatbot, si ottiene un'interfaccia interattiva che permette agli utenti di porre domande, esprimere pensieri o simulare vere e proprie conversazioni.

Questo approccio risulta particolarmente interessante se inserito nell'ambito della salute mentale, perché rende possibili servizi di supporto psicologico istantaneo e potenzialmente accessibili a un pubblico molto vasto, inclusi individui che per barriere economiche, culturali o geografiche non riescono a usufruire della terapia tradizionale[17].

Tuttavia, l'idea di utilizzare sistemi di intelligenza artificiale nel campo psicologico non è nuova: progetti come ELIZA (Joseph Weizenbaum, 1964-1966) aprirono la strada all'idea di un chatbot che "simulasse" un terapeuta rogersiano. Tuttavia, ELIZA si limitava a una riformulazione di quanto scritto

[15] Vedi E. C. Stade *et al.*, "Large Language Models Could Change the Future of Behavioral Healthcare: A Proposal for Responsible Development and Evaluation" in «Npj Mental Health Research», vol. III, 2024, n. 12, consultabile qui: <https://doi.org/10.1038/s44184-024-00056-z> (consultato il 24/04/25).

[16] Vedi K. Zhou *et al.*, "A survey of Large Language Model", in «arXiv», 2023, consultabile qui: <https://doi.org/10.48550/arXiv.2303.18223> (consultato il 24/04/25).

[17] Cfr. E. C. Stade *et al.,* cit.

dall'utente e non aveva alcuna comprensione semantica dei testi.

Gli attuali LLM, invece, vantano un'architettura che consente di elaborare in modo molto più sofisticato il contesto linguistico. Questa evoluzione è avvenuta grazie a reti neurali di tipo *transformer*, addestrate su *datasets* enormi. I risultati ottenuti mostrano come i modelli più recenti possano riprodurre stili comunicativi complessi, "simulare" empatia e persino adottare tecniche psicoterapeutiche specifiche[18].

Una delle maggiori promesse dei chatbot basati su LLM è la riduzione delle barriere di accesso alla cura psicologica. Alcuni studi sottolineano l'importanza di fornire un aiuto immediato a persone che, per vari motivi (distanze, difficoltà economiche, vergogna, stigma sociale), evitano di rivolgersi a un terapeuta umano. La possibilità di conversare con un chatbot in forma anonima, a qualsiasi ora del giorno, abbassa notevolmente la soglia di ingresso al percorso di sostegno emotivo[19].

Inoltre gli LLM possono alleggerire il carico di lavoro dei professionisti della salute mentale, occupandosi di compiti ripetitivi o di monitoraggio costante dei pazienti. Song *et al.* evidenziano come i chatbot possano raccogliere dati sulle condizioni emotive dell'utente, fornire test di screening preliminari e gestire esercizi cognitivi di routine[20]. I terapeuti, di conseguenza, possono dedicare più tempo alla relazione terapeutica vera e propria, analizzando in profondità le problematiche e intervenendo nei casi più complessi.

Sul piano psicoeducativo, i chatbot possono fornire informazioni su disturbi psicologici, suggerire strategie di auto-aiuto e invitare l'utente a riflettere su emozioni e comportamenti. A questo riguardo, lo studio di Nursen evidenzia come la possibilità di condividere il proprio vissuto in uno spazio percepito come sicuro aiuti i pazienti a esternare emozioni difficilmente ver-

[18] Vedi M. Xiao *et al.*, "HealMe: Harnessing Cognitive Reframing in Large Language Models for Psychotherapy" in «Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics» (vol. I: Long Papers), Bangkok 2024, pp. 1707-1725, consultabile qui: <https://doi.org/10.18653/v1/2024.acl-long.93> (consultato il 24/04/25).

[19] Cfr. E. C. Stade *et al.*, cit.

[20] Cfr. I. Song *et al.*, "The Typing Cure: Experiences with Large Language Model Chatbots for Mental Health Support" in «arXiv», 2024, consultabile qui: <https://doi.org/10.48550/ARXIV.2401.14362> (consultato il 24/05/25).

balizzabili in una seduta *vis-à-vis*[21]. La natura interattiva del chatbot incentiva, inoltre, l'autoesplorazione, elemento essenziale in qualsiasi percorso di salute mentale. Proprio per questo, una branca emergente della ricerca concerne l'addestramento degli LLM alle strategie di psicoterapia, al fine di migliorare l'integrazione e l'armonizzazione tra diversi approcci e metodologie. Ad esempio, Sun *et al.* propongono metodi per "istruire" gli LLM nel seguire protocolli terapeutici, come il *Motivational Interviewing* o la *Cognitive Behavioral Therapy* (CBT)[22]. In modo simile, Xiao *et al.* hanno dimostrato che un LLM appositamente addestrato per la ristrutturazione cognitiva può aiutare le persone a individuare e ridefinire pensieri negativi, migliorando la consapevolezza delle distorsioni cognitive[23].

I chatbot terapeutici basati su modelli di linguaggio di grandi dimensioni offrono quindi numerose opportunità nel contesto del supporto psicologico, pur presentando importanti limiti e problematiche da tenere in considerazione. Un primo vantaggio significativo è l'immediatezza e la continuità del sostegno offerto: gli utenti possono ricevere aiuto 24 ore su 24, elemento fondamentale specialmente per chi affronta episodi acuti di ansia o depressione in momenti della giornata in cui sarebbe difficile contattare uno specialista. A questo si aggiungono benefici economici e logistici, poiché i chatbot possono abbattere costi e distanze geografiche, ampliando notevolmente l'accessibilità alle cure psicologiche[24]. Inoltre, grazie alla loro flessibilità, i chatbot possono fornire strategie e suggerimenti personalizzati, adattandosi alle esigenze individuali degli utenti e favorendo una maggiore aderenza a pratiche di autoaiuto, come tecniche di *mindfulness* o esercizi cognitivi mirati. Un ulteriore vantaggio, messo in luce da Song *et al.*[25], riguarda il minor timore del giudizio: interagire con un'entità artificiale favorisce l'apertura emotiva e la condivi-

[21] K. Nurser, *"A Qualitative Exploration of Telling My Story in Mental Health Recovery"*, University of East Anglia, Norwich 2017, consultabile qui: <https://www.semanticscholar.org/paper/A-qualitative-exploration-of-Telling-My-Story-in-Nurser/6133508380a13be9c4af199770302335843e1e7f> (consultato il 24/04/25).

[22] X. Sun *et al.*, "Script-Strategy Aligned Generation: Aligning LLMs with Expert-Crafted Dialogue Scripts and Therapeutic Strategies for Psychotherapy", in «arXiv», 2024, consultabile qui: <https://doi.org/10.48550/ARXIV.2411.06723> (consultato il 24/05/25).

[23] M. Xiao *et al.*, cit.

[24] Cfr. E. C. Stade *et al.*, cit.

[25] I. Song *et al.*, cit.

sione di esperienze intime difficilmente esprimibili in presenza di un interlocutore umano.

Tuttavia, nonostante questi benefici, esistono alcune criticità che rendono necessario un approccio cauto all'uso di questi strumenti. Innanzitutto, gli LLM mancano di empatia autentica: come osservano Sun *et al.*, sebbene questi sistemi possano simulare un ascolto attivo, non sono in grado di sviluppare quella genuina sintonia emotiva che è alla base di un'efficace alleanza terapeutica[26]. Un secondo rischio riguarda i *bias* e i pregiudizi intrinseci ai dati utilizzati per addestrare questi modelli, problema già sollevato da Stade *et al.*, che raccomandano controlli rigorosi per evitare risposte discriminatorie o culturalmente inappropriate. La questione etica e della responsabilità clinica rappresenta un'ulteriore complessità: non è ancora chiaro come ripartire la responsabilità legale tra sviluppatori, professionisti sanitari e istituzioni in caso di errori o danni al paziente, soprattutto considerando la delicata gestione della privacy relativa ai dati sensibili. Un altro rischio importante è quello della dipendenza emotiva che può emergere dall'interazione costante con il chatbot, isolando l'utente e compromettendo il recupero, specialmente in disturbi in cui la dimensione relazionale risulta cruciale[27]. Infine, rimangono evidenti limiti tecnici nell'interazione a lungo termine: i modelli attuali possono perdere coerenza narrativa e faticare nel mantenere un'efficace continuità terapeutica nel corso di conversazioni prolungate, rendendo difficile il supporto in percorsi psicoterapeutici estesi e approfonditi[28]. Pertanto, l'integrazione equilibrata degli LLM nella pratica clinica dovrà necessariamente tenere conto tanto delle loro potenzialità quanto delle loro significative limitazioni.

L'uso crescente di chatbot basati su LLM nell'ambito della salute mentale richiederà in futuro una precisa regolamentazione e l'elaborazione di linee guida etiche da parte di organizzazioni internazionali, enti di ricerca e ordini professionali. Tali normative dovranno specificare chiaramente le competenze necessarie, i limiti di impiego e le modalità di supervisione da parte di specialisti qualificati. Parallelamente, sarà cruciale sviluppare approcci terapeutici ibridi, in cui l'intelligenza artificiale offra un primo livello di assistenza e mo-

[26] X. Sun *et al.*, op. cit.
[27] Vedi X. Sun *et al.*, op. cit.
[28] Vedi M. Xiao *et al.*, op. cit.

nitoraggio continuo, lasciando ai terapeuti umani il compito di gestire situazioni più complesse e garantire un intervento personalizzato e relazionale. Affinché tali strumenti siano davvero efficaci e affidabili, sarà inoltre necessario investire nella qualità dei dati utilizzati per l'addestramento dei modelli linguistici: occorrerà raccogliere *datasets* equilibrati, inclusivi e diversificati, rappresentativi di differenti realtà culturali e socio-economiche, con particolare attenzione alla privacy e alla gestione dei dati sensibili, in conformità con normative come il *Regolamento Generale sulla Protezione dei Dati* (GDPR), la normativa dell'Unione Europea entrata in vigore nel 2018 per regolamentare il trattamento dei dati personali e la loro circolazione, e l'*Health Insurance Portability and Accountability Act* (HIPAA), la legge statunitense che stabilisce standard per la protezione dei dati sensibili relativi alla salute dei pazienti e garantisce la portabilità dell'assicurazione sanitaria. Fondamentale, inoltre, sarà condurre valutazioni cliniche rigorose e a lungo termine, attraverso studi controllati e randomizzati che confrontino chiaramente l'efficacia dei chatbot terapeutici con quella delle terapie tradizionali o di altre forme di telepsicologia. Infine, una condizione imprescindibile per la realizzazione di tali prospettive sarà rappresentata dalla formazione dei professionisti della salute mentale, che dovranno acquisire competenze specifiche per integrare efficacemente le informazioni fornite dai chatbot nel loro lavoro quotidiano, riconoscendo con precisione quando sia necessario un intervento umano diretto e più avanzato.

Vulnerabilità e intelligenza artificiale: un nuovo paradigma terapeutico?

L'ingresso degli strumenti basati sull'intelligenza artificiale e, nello specifico, dei modelli di LLM nella sfera della salute mentale, apre una serie di riflessioni etiche e filosofiche che vanno ben oltre la semplice valutazione di efficacia terapeutica o l'identificazione di potenziali rischi tecnici. È in gioco, infatti, qualcosa di molto più complesso e profondamente radicato nell'esperienza umana: la vulnerabilità intesa non soltanto come condizione individualizzata di penuria e fragilità dei soggetti coinvolti, ma qualcosa di molto più vicino all'idea di una "intelligenza collettiva" di Pierre Lévy[29].

[29] P. Lévy, *L'intelligenza Collettiva. Per Un'antropologia Del Cyberspazio*, Feltrinelli, Milano 2002.

Dal momento che ciò che intendiamo per "umano" è sempre più il risultato di una sintesi costruita storicamente tra elementi simbolici, linguistici e tecnici – che contribuiscono attivamente a definirne l'identità ontologica – anche il concetto di vulnerabilità si trasforma: non più soltanto fragilità biologica o dipendenza sociale, ma un'esposizione strutturale che si manifesta all'incrocio tra natura, cultura e tecnologia. Essa cioè non è più solo né biologica (l'umano è fragile per costituzione naturale), né sociale (sono le istituzioni e i sistemi sociali che ci diamo che ci rendono esili). Non può essere ridotta a una condizione di passività o di fragilità da proteggere o riparare. Ma nemmeno può rimanere una mera lente concettuale tramite cui leggere antropologicamente l'evoluzione tecnologizzata – e tecnologizzante – della condizione umana. A contribuire a demolire queste passate visioni, da più parti, nel dibattito filosofico globale, sono diversi i segni che emergono e chiedono di ripensare la vulnerabilità in una chiave *differente*: non solo più interpretativa e comprensiva, cioè capace di allargare il campo e ospitare le nuove narrazioni della fragilità e della sofferenza umana provenienti dall'esperienza del mondo, ma una vulnerabilità al contempo *esperibile* e *immaginabile*, un'occasione epistemologica e normativa che al posto di contrapporsi a qualcosa di esteriore (vulnerabilità *vs* natura; vulnerabilità *vs* tecnologia), ne costituisce un attraversamento critico. Siamo "tecno-vulnerabili"[30]. Nell'essere vulnerabili siamo anche una messa in questione che rende la nostra condizione un'apertura dinamica e conoscitiva[31].

Cos'è una protesi robotica? Un ausilio per disabili che ripara o ripristina il corpo "normale"? Oppure un mezzo di potenziamento dell'umano oltre i propri limiti? Ci rende più umani o più cyborg? O forse entrambi? Dal dibattito sul "potenziamento umano" (*human enhancement*) a quello sulle differenze tra post- e transumanesimo[32], aprendo alle opportune incursioni degli studi di

[30] Vedi A. Carnevale, *Tecno-Vulnerabili. Per Un'etica Della Sostenibilità Tecnologica*, Orthotes, Napoli 2017. Si veda anche M. Coeckelbergh, *Human Being@ Risk: Enhancement, Technology, and the Evaluation of Vulnerability Transformations*, vol. I, Springer, Dordrecht and New York 2013.

[31] Vedi A. Carnevale, "Empowering Vulnerability: Decolonizing AI Ethics for Inclusive Epistemological Innovation", in «BioLaw Journal - Rivista di BioDiritto», 2024, pp. 25-37, consultabile qui: <https://doi.org/10.15168/2284-4503-3299> (consultato il 24/04/25).

[32] *Transumanesimo* e *post-umanesimo* sono due approcci distinti – e in parte conflittuali – rispetto alla trasformazione della condizione umana attraverso la tecnologia. Il transumanesimo è

genere[33] e dei *disability studies*[34], in controluce emerge, in tutto ciò, una generale disposizione dialettica ai temi della salute e della vulnerabilità, l'evidenza cioè che siamo perennemente corredati da dispositivi che sono sia un veleno sia una cura. Siamo cioè immersi nell'ambito del *pharmakon*, come Jacques Derrida aveva teorizzato nella sua *Farmacia di Platone*[35]. Il veleno non è qualcosa di esteriore, che ci colpisce e ci intossica allorché viene portato all'interno. Esso è un esterno che è sempre già presente all'interno, opera all'interno, ma non come mera presenza ma come sfondo, come traccia, come

---

un movimento filosofico e culturale che promuove il miglioramento delle capacità fisiche e cognitive dell'essere umano mediante l'uso di tecnologie avanzate, con l'obiettivo di superare i limiti biologici, la malattia e persino la morte. Il post-umanesimo, invece, è un paradigma critico che decostruisce l'idea di "umano" come soggetto centrale e autonomo della modernità occidentale. Esso rifiuta l'antropocentrismo e mette in discussione la distinzione netta tra umano, macchina e animale, privilegiando visioni relazionali, ibride e situate. Mentre il transumanesimo punta al potenziamento individuale, il post-umanesimo si concentra sulle interdipendenze materiali, simboliche e ambientali che costituiscono il vivente. Per un inquadramento generale, cfr. N. Bostrom, "Human Genetic Enhancements: A Transhumanist Perspective", in «The Journal of Value Inquiry», vol. XXXVII, 2003, n. 4, pp. 493-506, consultabile qui: <https://doi.org/10.1023/B:INQU.0000019037.67783.d5> (consultato il 24/04/25); F. Fukuyama, *Our Posthuman Future: Consequences of the Biotechnology Revolution*, Farrar, Straus and Giroux, New York 2003; N. Agar, *Humanity's End: Why We Should Reject Radical Enhancement*, MIT Press, Cambridge (MA) 2010; F. Battaglia e A. Carnevale, "Epistemological and Moral Problems with Human Enhancement", in «Humana. Mente Journal of Philosophical Studies», vol. VII, 2014, n. 26, pp. III–XX.

[33] Cfr. D. Haraway, *Manifesto cyborg. Donne, tecnologie e biopolitiche del corpo*, Feltrinelli, Milano 1995; S. Turkle, *La vita sullo schermo. Nuove identità e relazioni sociali nell'epoca di Internet*, Apogeo, Milano 2002; S. Plant, *Zeros and Ones: Digital Women and the New Technoculture*, Fourth Estate, London 1997; N. Katherine Hayles, *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*, University of Chicago Press, Chicago 1999; J. Wajcman, *TechnoFeminism*, Polity Press, Medford (MA) 2004.

[34] I *Disability Studies* sono un campo interdisciplinare che studia la disabilità non come una mera condizione medica o deficit individuale, ma come una costruzione sociale, culturale e politica. Questa prospettiva critica mette in discussione le narrazioni dominanti che associano la disabilità alla mancanza o all'anomalia, ponendo invece l'accento sulle barriere ambientali, istituzionali e comunicative che producono esclusione. Per una ricognizione generale sul tema, rimandiamo a titolo esemplificativo a N. Watson, A. Roulstone e C. Thomas (a cura di), *Routledge Handbook of Disability Studies*, 2nd Edition, Routledge, New York 2022; T. Shakespeare, *Disabilità e Società: Diritti, Falsi Miti, Percezioni Sociali*, Il Margine, Trento 2024.

[35] J. Derrida, *La Farmacia di Platone*, Jaca Book, Milano 2007.

allusione, aporia, come capro espiatorio. A tal proposito, Bernard Stiegler[36] – sulla scia di Derrida – ha sostenuto che l'uso dei mezzi tecnologici digitali non può mai essere ridotto esclusivamente a una funzione positiva (cura) o negativa (veleno), ma è sempre entrambe le cose allo stesso tempo. In questo senso, il *pharmakon* diventa un simbolo chiave di una possibile revisione della vulnerabilità in un'epoca di umanità digitale.

Una tale visione rappresenta un passaggio decisivo per comprendere e valutare pienamente il ruolo futuro delle tecnologie digitali nella salute mentale. Se la vulnerabilità non può più essere riferita a qualcosa di esterno – una situazione o un evento verso cui si è esposti e che potenzialmente ci influenza – essa implica non solo un'esposizione passiva alle circostanze, ma anche una disposizione relazionale che, in certi contesti, può essere potenziata dall'intervento tecnologico. È proprio in questa direzione che vanno interpretati i *potenziali* benefici riconosciuti all'utilizzo degli LLM in psicoterapia: questi strumenti permettono infatti di intervenire tempestivamente, con continuità e senza vincoli di spazio e tempo, facilitando un primo livello di assistenza immediata, spesso decisivo per soggetti che si trovano in condizioni di sofferenza acuta o emergenziale. Tale caratteristica rende il chatbot un mezzo prezioso per intervenire tempestivamente nei momenti di maggiore fragilità emotiva, specialmente per coloro che vivono in contesti isolati o che si trovano impossibilitati ad accedere facilmente ai servizi di cura tradizionali.

Tuttavia, questa potenzialità di intervento immediato e costante offerta dai chatbot terapeutici basati su LLM porta con sé anche una serie di rischi etici considerevoli, che vanno necessariamente considerati per prevenire effetti indesiderati e talvolta pericolosi. Innanzitutto, occorre tenere presente che la simulazione di empatia, seppur avanzata, non coincide con un'empatia autentica. Come dimostrato in alcuni degli studi menzionati *supra*[37], la capacità di ascolto e di interazione realistica degli LLM non può sostituire la relazione empatica che si instaura tra due esseri umani, e che rappresenta uno dei pilastri fondanti dell'efficacia terapeutica. L'assenza di questa sintonia autentica rischia di ridurre il valore terapeutico degli interventi, limitando così il

[36] B. Stiegler, *The Age of Disruption: Technology and Madness in Computational Capitalism*, Polity Press, Medford (MA) 2019.
[37] M. Xiao *et al.*, op. cit.; E. C. Stade *et al.*, op. cit.

loro impatto reale, specialmente nei casi più delicati e complessi.

Un ulteriore rischio che merita attenzione è quello dell'*autoterapia*. L'uso dei chatbot terapeutici basati su LLM potrebbe infatti indurre gli utenti a percepirli come strumenti autosufficienti per la gestione delle proprie emozioni, favorendo un rapporto solipsistico con la propria sofferenza piuttosto che un'elaborazione autentica e trasformativa della stessa. L'apparente disponibilità costante di un interlocutore virtuale, capace di rispondere immediatamente e con una parvenza di empatia, potrebbe generare l'illusione di un supporto terapeutico efficace, quando in realtà ciò che si ottiene è una conferma delle proprie convinzioni e stati emotivi, senza alcun processo di messa in discussione critica. Il problema non risiede tanto nel fatto che le persone gestiscano le proprie emozioni in uno spazio privato – cosa che, anzi, può essere segno di autonomia e maturità emotiva – quanto piuttosto nel rischio che il chatbot funzioni come *uno specchio che riflette e rafforza le alienazioni dell'utente* invece di offrirgli uno spazio per metterle in questione. A differenza di un approccio psicoterapeutico ibrido – in cui l'ausilio digitale potrebbe configurarsi come parte del setting e quindi aiutare a individuare contraddizioni, resistenze e schemi disfunzionali nel discorso del paziente – un chatbot privato opera secondo schemi predefiniti e tende a restituire risposte coerenti con il linguaggio e le emozioni espresse dall'utente, senza la capacità di contestualizzare criticamente il vissuto soggettivo.

In questo senso, il chatbot terapeutico potrebbe *cristallizzare certe forme di sofferenza piuttosto che favorirne il superamento*, poiché manca quella dialettica essenziale al processo terapeutico umano, in cui la crescita nasce spesso dalla frustrazione produttiva di essere messi di fronte ai propri limiti cognitivi ed emotivi. La terapia, infatti, non è solo un processo di ascolto, ma anche di sfida e decostruzione delle proprie narrazioni interiori.

Un altro aspetto critico riguarda il rischio che l'accessibilità dei chatbot terapeutici porti i gruppi sociali svantaggiati e subalterni a evitare completamente il confronto con professionisti umani. Se uno strumento digitale appare sempre disponibile e privo di costi emotivi ed economici, la tentazione di affidarsi esclusivamente ad esso può risultare forte, specialmente in contesti sociali in cui il disagio psicologico è stigmatizzato o dove l'accesso ai servizi di salute mentale è limitato da ragioni economiche.

Si potrebbe dunque assistere a una duplicazione dell'alienazione: alla *solitudine terapeutica* che si è visto prima, in cui il soggetto si chiude in una bolla

emotiva autoconfermativa, si aggiungerebbe anche una *povertà terapeutica*, una sorta di illusione di stare lavorando sulle proprie difficoltà quando in realtà si sta solo creando un ambiente di auto-rassicurazione che esonera dal mettersi davvero in discussione. In questo senso, il chatbot non solo non cura, ma rischia di diventare una tecnologia di *cristallizzazione del malessere*, una sorta di specchio che riflette il disagio senza offrire vie di uscita trasformative, quelle stesse che ci aiuterebbero a mettere a fuoco la base (di ingiustizia) sociale che spesso si nasconde dietro il disagio mentale. Una sorta di *dominio dell'esteriorità* a causa del quale la nostra coscienza storica sarebbe condannata a permanere nei suoi simulacri, non riuscendo a fare più piena esperienza della propria interiorità[38].

Da questo punto di vista, appare chiaro che l'integrazione degli strumenti tecnologici debba avvenire sempre in una prospettiva *ibrida*, dove la relazione umana rimane certo importante e significativa, ma senza per forza presupporre un qualche modello etico o normativo di soggettività. Pensiamo cioè che l'uso dell'IA nelle neuroscienze e nella psicoterapia contribuisca a far emergere una visione più realistica dell'IA come sistema di per sé né intelligente, né artificiale[39], bensì come supporto complementare di tipo *sociotecnico*[40] e non sostitutivo al lavoro clinico umano. Tale approccio, oltre a garantire di armonizzare sicurezza e innovazione terapeutica, permette di gestire situazioni che richiedono una sensibilità etica e clinica che gli algoritmi da soli attualmente non possono offrire. Ciò non significa, naturalmente, ridurre il ruolo degli strumenti digitali, ma implica la necessità di una governance precisa che definisca chiaramente i limiti e le condizioni di utilizzo di tali tecnologie, stabilendo standard rigorosi e promuovendo una continua supervisione professionale.

Tra le tante questioni, certamente quella dei *bias* e delle discriminazioni

---

[38] Cfr. R. Finelli e M. Gatto, *Il dominio dell'esteriore. Filosofia e critica della catastrofe*, Rogas, Roma 2024.

[39] K. Crawford, *Né artificiale né intelligente. Il lato oscuro dell'IA*, Il Mulino, Milano 2021.

[40] Sull'IA come sistema sociotecnico, si veda B. Green, "The Contestation of Tech Ethics: A Sociotechnical Approach to Technology Ethics in Practice", in «Journal of Social Computing», vol. II, n. 3, 2021, 209-225, consultabile qui: <https://doi.org/10.23919/JSC.2021.0018> (consultato il 24/04/25); A. Carnevale *et. al.*, "A Human-Centred Approach to Symbiotic AI", cit.

algoritmiche ancora rappresenta un'altra grande sfida etica[41]. Gli LLM sono spesso addestrati su *datasets* enormi che riflettono inevitabilmente diseguaglianze sociali, culturali e storiche. Come suggerito da Stade *et al.*[42], questi *bias*, se non correttamente affrontati in fase di progettazione e addestramento, possono portare a forme nuove di marginalizzazione e di stigmatizzazione, particolarmente problematiche nel caso della salute mentale, dove la vulnerabilità degli utenti amplifica il rischio di subire danni significativi da comportamenti discriminatori del sistema.

Tuttavia, l'etica dell'IA da sola potrebbe non bastare[43]. Ciò che è considerato bene o è male dipende da numerosissime variabili, molte delle quali sono affette da logiche di potere. Esiste infatti il rischio di una "colonizzazione digitale" e di uno sfruttamento delle vulnerabilità per ragioni economiche o di mercato, e ciò richiama l'attenzione sulla necessità di introdurre una prospettiva decoloniale che consideri le strutture di potere economico e culturale alla base della tecnologia stessa[44]. In altre parole, l'etica applicata all'IA deve includere riflessioni critiche sui rapporti di potere, sui rischi di appropriazione e sfruttamento dei dati personali e sulle logiche di mercato che orientano l'uso delle tecnologie[45]. Occorre, quindi, porre attenzione non solo agli effetti tecnici dell'IA ma anche alle condizioni sociali e politiche entro cui queste tec-

---

[41] Vedi T. Numerico, *Big Data e Algoritmi*, op. cit.; L. Floridi, *Etica Dell'intelligenza Artificiale: Sviluppi, Opportunità, Sfide*, Raffaello Cortina, Milano 2022.

[42] E. C. Stade *et al.*, cit.

[43] Cfr. B. Mittelstadt, "Principles Alone Cannot Guarantee Ethical AI", in «*Nature Machine Intelligence*», vol. I, 2019, n. 11, 2019, pp. 501-507, consultabile qui: <https://doi.org/10.1038/s42256-019-0114-4> (consultato il 24/04/25); L. Munn, "The Uselessness of AI Ethics", in «*AI and Ethics*», vol. III, 2023, n. 3, 869-877, consultabile qui: <https://doi.org/10.1007/s43681-022-00209-w> (consultato il 24/04/25).

[44] Cfr. N. Couldry e U. A Mejias, "Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject", in «Television & New Media», vol. XX, 2019, n. 4, pp. 336-349, consultabile qui: <https://doi.org/10.1177/15274764187966> (consultato il 24/04/25); S. Mohamed *et al.*, "Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence", in «Philosophy & Technology», vol. XXXIII, 2020, n. 4, pp. 659-684, consultabile qui: <https://doi.org/10.1007/s13347-020-00405-8> (consultato il 24/04/25); P. Ricaurte, "Data Epistemologies, the Coloniality of Power, and Resistance", in «Television & New Media», vol. XX, 2019, n. 4, pp. 350-365, consultabile qui: <https://doi.org/10.1177/1527476419831640> (consultato il 24/04/25); A. Carnevale, "Empowering Vulnerability", cit.

[45] Cfr. K. Crawford, *Né artificiale né intelligente*, op. cit.; T. Numerico, *Big Data e Algoritmi*, cit.

nologie si collocano, sviluppando un'etica della tecnologia che tenga conto della complessità sociale e culturale degli utenti, e che sia consapevole delle implicazioni più ampie e strutturali della digitalizzazione, incluso un ripensamento del corpo oltre ogni ontologico binarismo[46].

La sfida etica posta dall'integrazione di tecnologie digitali nella salute mentale consiste, quindi, nel riconoscere e promuovere questo rapporto co-creativo tra tecnologia, vulnerabilità ed esperienza umana, sapendo bilanciare le potenzialità degli strumenti digitali con le loro intrinseche limitazioni, incluse le forme di *solitudine* e *povertà terapeutica* che abbiamo descritto. Soltanto così potrà emergere un'alleanza terapeutica al passo coi tempi, che armonizzi efficacemente le opportunità della tecnologia digitale con la capacità insostituibile della relazione umana di rispondere alle esigenze di persone che si *sentono* vulnerabili, trasformando la loro fragilità, percepita o reale che sia, in occasione di conoscenza, emancipazione e crescita individuale e collettiva.

Di tale sfida, a conclusione di questo nostro intervento, vorremmo mettere in luce un ultimo aspetto correlato. Certamente, l'uso dei *Large Language Models* nelle neuroscienze e nella psicoterapia impone la necessità di formare professionisti sociosanitari capaci di gestire consapevolmente i benefici della rivoluzione digitale, mettendoli al servizio della salute mentale dei pazienti. Ma a nostro avviso, tale esigenza non riguarda solo i professionisti della cura: è l'intera *catena sociotecnica di valore*[47] che deve essere oggetto di attenzione etica. Con questa espressione intendiamo l'insieme articolato di attori, strumenti, norme, dati e pratiche che, in modo interdipendente, partecipano alla progettazione, implementazione e uso dei sistemi di IA. In questo contesto, è fondamentale definire con chiarezza ruoli e responsabilità lungo l'intero ciclo di vita della tecnologia, dal design alla regolazione, fino all'interazione con

[46] Vedi F. R. Recchia Luciani, "Binarismo ontologico e 'differenza sessuale': il protofemminismo, il femminismo storico e la difficile liberazione delle donne", in *La memoria nella costruzione dell'esperienza. Dalla storia della filosofia alle scienze cognitive*, R. Roni (a cura di), Maria Pacini Fazzi, Lucca 2024, pp. 91-104.

[47] Per un approfondimento sul significato e funzione della "catena di valore" (*value chain*) in senso sociotecnico, rimandiamo a un saggio di Luciano Floridi in cui l'autore discute la responsabilità diffusa nelle filiere sociotecniche digitali, sottolineando come l'etica non possa limitarsi al momento d'uso ma debba estendersi a tutta la catena: L. Floridi, "Soft Ethics and the Governance of the Digital", in «Philosophy & Technolology», vol. XXXI, 2018, pp. 1-8, consultabile qui: <https://doi.org/10.1007/s13347-018-0303-9> (consultato il 24/04/25).

l'utente finale. E dunque accanto allo psicologo, al terapeuta, al medico, occorrono esperti di etica. L'adozione indiscriminata di principi generali come trasparenza, responsabilità e inclusività rimane insufficiente, se non affiancata da un chiaro quadro etico-operativo che comprenda strumenti specifici per identificare rischi etici latenti nell'utilizzo dei dati e degli algoritmi. È necessario, pertanto, un approccio etico più focalizzato sui dettagli concreti della gestione dei dati e sulle pratiche algoritmiche, una "infraetica" che renda operativi i valori morali astratti[48]. Una simile figura, qualora inserita proattivamente nella catena di valore, aiuterebbe a mettere in guardia dalla cosiddetta *ethics washing*[49], quel processo per cui le Big Tech sono tentate *by-default* di sfruttare retoricamente l'etica per nascondere logiche di profitto spesso incompatibili con un benessere umano e sociale realmente esperibile e alla portata di tutti e tutte.

Conclusioni

L'uso dei *Large Language Models* nella psicoterapia rappresenta una frontiera in continua evoluzione, capace di ridefinire il rapporto tra tecnologia e benessere mentale. Se da un lato l'accessibilità, l'immediatezza e il supporto costante offerti dai chatbot terapeutici aprono nuove prospettive di intervento, dall'altro emergono interrogativi cruciali legati all'assenza di empatia autentica, ai *bias* algoritmici e ai limiti strutturali della relazione terapeutica mediata dall'intelligenza artificiale.

L'analisi condotta in questo lavoro suggerisce che il valore degli LLM nel contesto della salute mentale non risiede nella loro capacità di sostituire il terapeuta umano, bensì nel loro potenziale di supporto e integrazione all'interno di modelli ibridi di assistenza psicologica. Tuttavia, il fatto che alcuni utenti si rivolgano a questi strumenti con aspettative di sollievo o ascolto apre una domanda fondamentale: la cura risiede esclusivamente nella relazio-

[48] Cfr. E. Panai, "The Latent Space of Data Ethics", in «AI & SOCIETY», vol. XXXIX, 2024, n. 6, pp. 2647-2665, consultabile qui: <https://doi.org/10.1007/s00146-023-01757-3> (consultato il 24/04/25).

[49] Cfr. J. Steinhoff, "AI Ethics as Subordinated Innovation Network", in «AI & SOCIETY», vol. XXXIX, 2024, n. 4, 2024, pp. 1995-2007, consultabile qui: <https://doi.org/10.1007/s00146-023-01658-5> (consultato il 24/04/25).

ne interpersonale, oppure può attivarsi anche attraverso una simulazione credibile di dialogo, che genera anche solo una sensazione – pur illusoria – di riconoscimento?

La psicoterapia, nella sua essenza, resta un processo relazionale e trasformativo. Se l'interazione linguistica con un chatbot è in grado di produrre un effetto soggettivo di conforto, ciò interroga profondamente le nostre concezioni sulla natura della cura: è sufficiente la forma dialogica per attivare dinamiche terapeutiche, o è l'intenzionalità dell'altro umano a fondare il processo di guarigione?

A queste domande non si può rispondere in modo definitivo, ma una cosa è chiara: l'integrazione dell'IA nella salute mentale richiede regolamentazione attenta, supervisione etica rigorosa e una riflessione continua sugli effetti a lungo termine. Non si tratta di opporre umanità e tecnologia, ma di comprendere in che misura l'adozione di strumenti digitali possa arricchire, affiancare o eventualmente distorcere il senso della relazione terapeutica[50].

[50] Riconoscimenti: This work was partially supported by the project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

BIBLIOGRAFIA

AGAR N., *Humanity's End: Why We Should Reject Radical Enhancement*, MIT Press, Cambridge (MA) 2010.

ALIM E. *et al.*, "Enteric Nervous System, Gut-Brain Connection and Related Neurodevelopmental Disorders", in «Anatomy», vol. XIV, 2020, n. 1, pp. 61-67, consultabile qui: <https://doi.org/10.2399/ana.20.008> (consultato il 24/04/25).

BATTAGLIA F. e CARNEVALE A., "Epistemological and Moral Problems with Human Enhancement", in «Humana. Mente Journal of Philosophical Studies», vol. VII, 2014, n. 26, pp. III-XX.

BEAVER I., "*Is AI at Human Parity Yet? A Case Study on Speech Recognition*", in «AI Magazine», vol. LXIII, 2022, n. 4, pp. 386-389.

BOSTROM N., "Human Genetic Enhancements: A Transhumanist Perspective", in «The Journal of Value Inquiry», vol. XXXVII, 2003, n. 4, pp. 493-506, consultabile qui: <https://doi.org/10.1023/B:INQU.0000019037.67783.d5> (consultato il 24/04/25).

CARNEVALE A. *et. al.*, "A Human-Centred Approach to Symbiotic AI: Questioning the Ethical and Conceptual Foundation", in «Intelligenza Artificiale», 2024, pp. 1-12, consultabile qui: <https://doi.org/10.3233/IA-240034> (consultato il 24/04/25).

CARNEVALE A., "Empowering Vulnerability: Decolonizing AI Ethics for Inclusive Epistemological Innovation", in «BioLaw Journal - Rivista di BioDiritto», 2024, pp. 25-37, consultabile qui: <https://doi.org/10.15168/2284-4503-3299> (consultato il 24/04/25).

—, *Tecno-Vulnerabili. Per Un'etica Della Sostenibilità Tecnologica*, Orthotes, Napoli 2017.

COECKELBERGH M., *Human Being@ Risk: Enhancement, Technology, and the Evaluation of Vulnerability Transformations*, vol. I, Springer, Dordrecht and New York 2013.

COULDRY N. e MEJIAS U. A., "Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject", in «Television & New Media», vol. XX, 2019, n. 4, pp. 336-349, consultabile qui: <https://doi.org/10.1177/15274764187966> (consultato il 24/04/25).

CRAWFORD K., *Né artificiale né intelligente. Il lato oscuro dell'IA*, Il Mulino, Bologna 2021.

CRISTIANINI N., *La Scorciatoia*, il Mulino, Bologna 2023.

CSÍKSZENTMIHÁLYI M., *Flow: The Psychology of Optimal Experience*, Harper & Row, New York 1990.

DAMASIO A., *L'errore di Cartesio: Emozione, ragione e cervello umano*, Adelphi, Milano 1994.

DENNETT D.C., *The Intentional Stance*, MIT Press, Cambridge (MA) 1987.

DERRIDA J., *La Farmacia di Platone*, Jaca Book, Milano 2007.

FINELLI R. e GATTO M., *Il dominio dell'esteriore. Filosofia e critica della catastrofe*, Rogas, Roma 2024.

FOSTER J. A. e MCVEY NEUFELD K.-A., "Gut–Brain Axis: How the Microbiome Influences Anxiety and Depression", in «Trends in Neurosciences», vol. XXXVI, 2013, n. 5, pp. 305-312, consultabile qui: <https://doi.org/10.1016/j.tins.2013.01.005> (consultato il 24/04/25).

FLORIDI L., *Etica Dell'intelligenza Artificiale: Sviluppi, Opportunità, Sfide*, Raffaello Cortina, Milano 2022.

FLORIDI L., "Soft Ethics and the Governance of the Digital", in «Philosophy & Technolology», vol. XXXI, 2018, pp. 1-8, consultabile qui: < https://doi.org/10.1007/s13347-018-0303-9> (consultato il 24/04/25).

FUKUYAMA F., *Our Posthuman Future: Consequences of the Biotechnology Revolution*, Farrar, Straus and Giroux, New York 2003.

GREEN B., "The Contestation of Tech Ethics: A Sociotechnical Approach to Technology Ethics in Practice", in «Journal of Social Computing», vol. II, 2021, n. 3, pp. 209-225, consultabile qui: <https://doi.org/10.23919/JSC.2021.0018> (consultato il 24/04/25).

HARAWAY D., "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective", in «Feminist Studies», vol. XIV, 1988, n. 3, pp. 575-599.

—, *Manifesto cyborg. Donne, tecnologie e biopolitiche del corpo*, Feltrinelli, Milano 1995.

HAYLES K. N., *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*, University of Chicago Press, Chicago 1999.

HEIDEGGER M., *La questione della tecnica* (1953), goWare, Firenze 2017.

HORGAN T. e TIENSON J., *Connectionism and the Philosophy of Psychology,* MIT Press, Cambridge (MA) 1996.

KAHNEMAN D., *Pensieri lenti e veloci*, Mondadori, Milano 2012.
LÉVY P., *L'intelligenza Collettiva. Per Un'antropologia Del Cyberspazio*, Feltrinelli, Milano 2002.
MITTELSTADT B., "Principles Alone Cannot Guarantee Ethical AI", in «Nature Machine Intelligence», vol. I, 2019, n. 11, pp. 501-507, consultabile qui <https://doi.org/10.1038/s42256-019-0114-4> (consultato il 24/04/25).
MOHAMED S. *et al.*, "Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence", in «Philosophy & Technology», vol. XXXIII, 2020, n. 4, pp. 659-684, consultabile qui: <https://doi.org/10.1007/s13347-020-00405-8> (consultato il 24/04/25).
MUNN L., "The Uselessness of AI Ethics", in «AI and Ethics», III, 2023, n. 3, pp. 869-877, consultabile qui: <https://doi.org/10.1007/s43681-022-00209-w> (consultato il 24/04/25).
NUMERICO T., *Big data e algoritmi. Prospettive critiche*, Carocci, Roma 2021.
NURSER K., "A Qualitative Exploration of Telling My Story in Mental Health Recovery", 2017, consultabile qui: <https://www.semanticscholar.org/paper/A-qualitative-exploration-of-Telling-My-Story-in-Nurser/6133508380a13be9c4af199770302335843e1e7f> (consultato il 24/04/25).
PANAI E., "The Latent Space of Data Ethics", in «AI & SOCIETY», vol. XXXIX, 2024, n. 6, pp. 2647-2665, consultabile qui: <https://doi.org/10.1007/s00146-023-01757-3> (consultato il 24/04/25).
PLANT S., *Zeros and Ones: Digital Women and the New Technoculture,* Fourth Estate, London 1997.
RECCHIA LUCIANI F. R., "Binarismo ontologico e 'differenza sessuale': il protofemminismo, il femminismo storico e la difficile liberazione delle donne", in *La memoria nella costruzione dell'esperienza. Dalla storia della filosofia alle scienze cognitive*, R. Roni (a cura di), Maria Pacini Fazzi, Lucca 2024, pp. 91-104.
RICAURTE P., "Data Epistemologies, the Coloniality of Power, and Resistance", in «Television & New Media», vol. XX, 2019, n. 4, pp. 350-365, consultabile qui: <https://doi.org/10.1177/1527476419831640> (consultato il 24/04/25).
ROSENBERGER R. e VERBEEK. P.-P. (a cura di), *Postphenomenological Investigations. Essays on Human-Technology Relations*, Lexington Books, London 2015.

SEARLE J., "Minds, Brains and Programs", in «The Behavioral and Brain Sciences», vol. III, 1980, n. 3, pp. 417-457.

SHAKESPEARE T., *Disabilità e Società: Diritti, Falsi Miti, Percezioni Sociali*, Il Margine, Trento 2024.

SHOHAM Y. *et al.*, *The AI Index 2018 Annual Report*, AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford (CA) 2018, consultabile qui: <https://hai.stanford.edu/ai-index/2018-ai-index-report> (consultato il 24/04/25).

SIMONDON G., *Du mode d'existence des objiets techniques*, Aubier-Montaigne, Paris 1958.

SONG I. *et al.*, "The Typing Cure: Experiences with Large Language Model Chatbots for Mental Health Support", in «arXiv», 2024, consultabile qui: https://doi.org/10.48550/ARXIV.2401.14362 (consultato il 24/04/25).

STADE E. C. *et al.*, "Large Language Models Could Change the Future of Behavioral Healthcare: A Proposal for Responsible Development and Evaluation", in «Npj Mental Health Research», vol. III, 2024, n. 12, consultabile qui: <https://doi.org/10.1038/s44184-024-00056-z> (consultato il 24/04/25).

STEINHOFF J., "AI Ethics as Subordinated Innovation Network", in «AI & SOCIETY», vol. XXXIX, 2024, n. 4, pp. 1995-2007, consultabile qui: <https://doi.org/10.1007/s00146-02301658-5> (consultato il 24/04/25).

STIEGLER B., *La technique et le temps, vol. I. La faute d'Épiméthée*, Galilée, Paris 1994.

—, *The Age of Disruption: Technology and Madness in Computational Capitalism*, Polity Press, Medford (MA) 2019.

SUN X. *et al.*, "Script-Strategy Aligned Generation: Aligning LLMs with Expert-Crafted Dialogue Scripts and Therapeutic Strategies for Psychotherapy", in «arXiv», 2024, consultabile qui: <https://doi.org/10.48550/ARXIV.2411.06723> (consultato il 24/04/25).

TURKLE S., *La vita sullo schermo. Nuove identità e relazioni sociali nell'epoca di Internet*, Apogeo, Milano 2002.

VERBEEK P.-P., *What Things Do. Philosophical Reflections on Technology, Agency, and Design*, Penn State University Press, University Park (PA) 2005.

WAJCMAN J., *TechnoFeminism*, Polity Press, Medford (MA) 2004.

WATSON N., ROULSTONE A. e THOMAS C. (a cura di), *Rout-ledge Handbook of Disability Studies*, 2nd Edition, Routledge, New York 2022.

XIAO M. *et al.*, "HealMe: Harnessing Cognitive Reframing in Large Language Models for Psychotherapy", in «Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (vol. I: Long Papers)», Bangkok 2024, pp. 1707-1725, consultabile qui: <https://doi.org/10.18653/v1/2024.acl-long.93> (consultato il 24/04/25).

ZHOU K. *et al.*, "A survey of Large Language Model", in «arXiv», 2023, consultabile qui: <https://doi.org/10.48550/arXiv.2303.18223> (consultato il 24/04/25).

SITOGRAFIA

"Mechanosensitive Neurons Innervating the Gut and Heart Control Metabolic and Emotional State", in «Nature Metabolism», vol. VII, 2025, pp. 249-250, consultabile qui: <https://doi.org/10.1038/s42255-024-01208-3> (consultato il 24/04/25).