SANITÀ DIGITALE E IA EMOTIVA ALLA PROVA DELL'EGUAGLIANZA: DAI *BIAS* ALLE PROSPETTIVE PER L'INCLUSIONE*

di Sena Santini**

Sommario. 1. L'impatto dirompente dei sistemi di intelligenza artificiale nel settore sanitario. – 2. Definizione e classificazione dei *bias*. – 3. *Bias* e discriminazione algoritmica in sanità. – 4. IA emotiva e rischi per i diritti fondamentali. – 5. IA emotiva applicata al settore sanitario e *bias*. – 6. Mitigazione dei *bias* e IA come strumento di democratizzazione dei processi decisionali. – 7. Sistemi di IA quale veicolo di inclusione sociale. – 8. Riflessioni conclusive.

1. L'impatto dirompente dei sistemi di intelligenza artificiale nel settore sanitario.

L'impatto pervasivo di tecnologie basate sull'IA sta progressivamente coinvolgendo anche il settore sanitario, dove i sistemi di IA sono attualmente impiegati prevalentemente come strumenti di supporto diagnostico[1]. Al contempo, si assiste a un'espansione significativa nel loro impiego per l'assistenza alla salute mentale, con prospettive che arrivano ormai a includere la realizzazione di piani terapeutici personalizzati e lo sviluppo di piattaforme terapeutiche virtuali[2].

Il recente dibattito scientifico in tema di problemi etici e giuridici connessi alle *AI based technologies* ha evidenziato che la presenza di *bias* sistemici di diversa natura può esacerbare le disuguaglianze preesistenti legate a fattori soggettivi quali *status* socio-economico, etnia, genere, condizioni di salute dei pazienti[3]. Nel settore della sanità, il rischio di esiti discriminatori legati alle decisioni algoritmiche risulta particolarmente rilevante, in quanto sono in gioco diritti fondamentali degli individui. Tale preoccupazione si affianca a *key-issue* di sempre maggior rilievo quali la trasformazione della relazione medico-paziente, il rischio di *deskilling* dei professionisti e la tutela del dato sanitario[4].

Come indicato da linee guida e documenti chiave, tra cui il rapporto *Guidance on Ethics and Governance of Artificial Intelligence for Health* (OMS, 2021)[5], il *Blueprint for an AI Bill of Rights*

---

* *Sottoposto a referaggio*.

** Dottoranda di ricerca in Diritto Costituzionale – Università di Parma.

[1] Per una disamina completa sul tema cfr. Consiglio Superiore di Sanità, *I sistemi di intelligenza artificiale come strumento di supporto alla diagnostica*, Ministero della Salute, 9 novembre 2021.

[2] D. Olawade, *Enhancing mental health with Artificial Intelligence: Current trends and future prospects*, in *Journal of Medicine, Surgery, and Public Health*, Vol. 3, 2024; C, Maurizio, *Chatbot terapeutici: rivoluzione o placebo per la salute mentale?,* in *Agenda Digitale*, 22 maggio 2025.

[3] Per una panoramica sul tema dei *bias* e discriminazione algoritmica si rimanda, senza pretesa di esaustività, a P. Traverso, *Breve introduzione tecnica all'Intelligenza Artificiale,* in *DPCE online*, 1, 2022; P. Zuddas, *Intelligenza artificiale e discriminazioni,* in *Consulta Online, Liber Amicorum per Pasquale Costanzo*, 2020; V. Molaschi, *Algoritmi e discriminazione,* in *Fundamental Rights,* 2, 2022, 19-29, *passim*.

[4] D. Morana, T. Balduzzi, F. Morganti, *La salute "intelligente": eHealth, consenso informato e principio di non-discriminazione,* in *Federalismi.it*, 34, 2022, 128 ss.

[5] OMS, *Ethics and governance of artificial intelligence for health: WHO guidance*, 2021.

(White House, 2022)[6] e *Artificial Intelligence in Healthcare* redatto dal Servizio Ricerca del Parlamento europeo (EPRS, 2022), la presenza di *bias* algoritmici nella sanità può condurre a valutazioni cliniche meno accurate e a una sottostima sistematica dei bisogni di cura e di assistenza, penalizzando in particolare gli strati della popolazione più vulnerabili[7].
Il presente contributo si pone l'obiettivo di analizzare le sfide e le opportunità poste dal crescente impiego di sistemi di IA nel settore sanitario, con un *focus* particolare sulle applicazioni di sistemi di *emotional AI*, categoria particolarmente esposta al rischio di *bias*, in ragione della natura intrinsecamente soggettiva e complessa del dato emotivo. A tal fine, si esamineranno casi concreti di esiti discriminatori derivanti da un utilizzo acritico dei sistemi di IA e si evidenzieranno strategie di mitigazione dei *bias*, volte a garantire la protezione delle minoranze e a promuovere un accesso equo e inclusivo ai servizi sanitari, nel rispetto dei principi di uguaglianza, trasparenza e non discriminazione, quali principi guida del Regolamento (UE) 2024/1689.

Anche alla luce della recente l. n. 132/2025, rubricata *Disposizioni e deleghe al Governo in materia di intelligenza artificiale,* che all'art. 7 enfatizza sia il diritto a un accesso non discriminatorio alle prestazioni sanitarie, sia la promozione dello sviluppo di tecnologie inclusive per persone con disabilità, si illustrerà come, accanto ai rischi, l'IA possa offrire opportunità per alcuni gruppi vulnerabili. Ciò, non solo attraverso soluzioni quali – a titolo meramente esemplificativo – sistemi di riconoscimento delle emozioni facciali, *carebots*[8] e assistenti vocali, che favoriscono l'accessibilità e il pieno sviluppo delle persone con disabilità, ma anche configurandosi, in molti casi, gli stessi sistemi di IA come un mezzo capace di ridurre i *bias* cognitivi e sociali e di democratizzare l'accesso all'assistenza sanitaria, nell'ottica di una concreta attuazione del principio di uguaglianza sostanziale.

In tale prospettiva, l'intelligenza artificiale non deve essere considerata esclusivamente quale fattore di rischio per i diritti fondamentali, ma anche come possibile strumento attraverso cui perseguire gli obiettivi di inclusione e di rimozione degli ostacoli che, ai sensi dell'art. 3, secondo comma, Cost., limitano di fatto la libertà e l'eguaglianza dei cittadini. L'uso delle tecnologie digitali in ambito sanitario assume, così, una dimensione non soltanto efficientistica, ma anche solidaristica e redistributiva.
Il presente contributo pone il *focus* sui *bias* legati a sistemi di IA, ma è opportuno sottolineare che anche strumenti di sanità digitale apparentemente neutri, come il Fascicolo Sanitario Elettronico[9], possono essere idonei a incorporare *bias* impliciti, rispecchiando discriminazioni esistenti[10].

***2. Definizione e classificazione dei* bias.** È rilevante notare che le interazioni degli utenti

[6] White House Office of Science and Technology Policy, *Blueprint for an AI Bill of Rights*, 2022.
[7] C. Nardocci, *Artificial Intelligence-based Discrimination: Theoretical and Normative Responses. Perspectives from Europe,* in *DPCE Online*, 3, 2023; G. Giacomini, C. Aprilis, *Digital Discrimination. The Challenge of Bias and Transparency in AI,* in *Lessico di Etica Pubblica*, 1, 2025.
[8] Per un approfondimento sui robot assistivi e il concetto di *carebots* o *care robots* si veda E. Ferioli, *Digitalizzazione, Intelligenza Artificiale e Robot nella tutela della salute,* in A. D'Aloia (a cura di), *Intelligenza artificiale e diritto. Come regolare un mondo nuovo,* Milano, 2020, 423-450, 426.
[9] Per un quadro normativo del FSE si rimanda a E. Catelani, *Nuove tecnologie e tutela del diritto della salute: potenzialità e limiti dell'uso della Blockchain,* in *Federalismi.it,* 4, 2022, 211-225.
[10] Un esempio in tal senso è lo studio condotto presso un ospedale universitario di Chicago, che ha evidenziato un uso significativamente più frequente di descrittori negativi nei fascicoli clinici di pazienti afroamericani, nella sezione *history and physical notes,* equivalente al profilo sanitario sintetico italiano, nonostante condizioni socio-demografiche e cliniche analoghe, cfr. M. Sun *et al.*, *Negative patient descriptors: Documenting racial bias in the Electonic Health Record*, in *Health Affairs*, 2, 2022, 203 ss.; D. Morana, T. Balduzzi, F. Morganti, *La salute "intelligente": eHealth, consenso informato e principio di non-discriminazione*, cit., 146.

con sistemi di IA, in particolare in settori come quello sanitario, sono attualmente segnate da una diffusa mancanza di fiducia. Tale sfiducia sembra derivare in parte dall'inconoscibilità dei processi interni (il c.d. *black box problem*[11]) e in parte dai potenziali esiti discriminatori derivanti da sistemi di IA affetti da *bias*.

Non a caso, tale preoccupazione viene fatta propria anche dal legislatore italiano che all'art. 7, cO. 2, l. n. 132/2025 specifica che «L'introduzione di sistemi di intelligenza artificiale nel sistema sanitario non può selezionare e condizionare l'accesso alle prestazioni sanitarie secondo criteri discriminatori».

Le fonti di *bias* che possono influenzare i sistemi di IA sono molteplici e incidono su diverse fasi dello sviluppo e dell'implementazione degli algoritmi[12]. Possiamo riassumerle in questa tripartizione.

I *bias* possono, in prima battuta, essere introdotti per mezzo del *dataset* con cui è addestrato il sistema – sia nella fase di raccolta dati, sia di etichettatura (*labeling*) dei dati stessi – possiamo parlare, dunque, di *bias* nel *dataset*. Le distorsioni possono, poi, emergere dalla progettazione, dal funzionamento e dalla selezione del sistema impiegato, in quanto l'uso di un algoritmo inadatto alla complessità o alla specificità di un dato compito può compromettere l'accuratezza degli *output*. Indipendentemente dalla presenza o assenza di *bias* nei dati, se si sceglie un algoritmo inadatto al tipo di problema, il modello farà previsioni imprecise. Per sintetizzare possiamo dunque parlare di *bias* nella scelta o nella progettazione dell'algoritmo.

La terza fonte di *bias* è, infine, rappresentata dall'elemento umano. È possibile, infatti, che inconsapevolmente vengano introdotti dagli sviluppatori *bias* impliciti, o influenze legate a esperienze, convinzioni o aspettative, che non risultano necessariamente valide o generalizzabili. Possiamo, dunque, parlare di *bias* introdotti dagli sviluppatori. Facendo seguito a quanto finora detto, senza pretesa di esaustività, si possono individuare alcune principali tipologie di *bias*[13]. Il *bias* algoritmico deriva da assunzioni errate nel processo di apprendimento automatico e può essere introdotto a causa del *dataset* o per scelte distorte nella progettazione stessa dell'algoritmo[14]. Per quanto attiene a distorsioni derivanti dal *dataset*, rilevano, poi, il *bias* da esclusione e il *bias* di selezione. Il *bias* da esclusione si verifica, in particolare, quando dati rilevanti vengono omessi dal *dataset*, spesso perché lo sviluppatore non ha identificato fattori nuovi o cruciali. Il *bias* di selezione (o di misurazione) si verifica, infine, quando il campione selezionato non riflette la popolazione che dovrebbe essere considerata. Il *bias* può essere, poi, di tipo cognitivo ed essere introdotto nell'algoritmo da chi cura l'addestramento e la valutazione del modello. Tale *bias* si verifica quando la tecnologia necessita dell'intervento umano, introducendo l'inevitabile fallibilità del giudizio umano. Strettamente connesso a quello cognitivo è, poi, il *bias* di conferma che si manifesta quando l'IA si basa su tendenze già presenti nei dati, confermando pregiudizi esistenti anziché individuare nuovi schemi o soluzioni[15].

---

[11] Per un'analisi del tema cfr. *ex multis,* G. Lo Sapio, *La black box: l'esplicabilità delle scelte algoritmiche quale garanzia di buona amministrazione,* in *Federalismi.it,* 16, 2021; *C.* Zednik, *Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence,* in *Philosophy & technology*, 34.2, 2021.

[12] K. Mavrogiorgos, A. Kiourtis, A. Mavrogiorgou, A. Menychtas, D., Kyriazis, *Bias in machine learning: A literature review,* in *Applied Sciences*, 14(19), 2024, 2.

[13] J. Holdsworth, *What is AI bias?,* in *IBM*, 12 Dicembre 2023.

[14] L. Di Giacomo, *Algoritmi e bias: come l'intelligenza artificiale può riprodurre o combattere i pregiudizi*, in *Diritto.it*, 23 agosto 2024.

[15] Si fa riferimento a tal proposito a *output* confermativi come quelli messi in luce dalla letteratura nel noto Caso Compas. Sul punto cfr. A. D'Aloia, *Intelligenza artificiale, società algoritmica, dimensione giuridica. Lavori in corso*, in *Quaderni Costituzionali*, 3, 2022, 667; S. Carrer, *Se l'amicus curiae è un algoritmo: il chiacchierato caso Loomis alla Corte Suprema del Wisconsin*, in *Giurisprudenza Penale Web,* 4, 2019; A. Simoncini, *L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà*, in *BioLaw Journal*, 1, 2019, *passim.*

Con particolare attenzione ai rischi posti dai sistemi *self-learning*, è inevitabile osservare l'impatto dello sviluppo dell'IA sul principio di eguaglianza, tanto in senso negativo – ad esempio, a scapito delle minoranze o mediante il rafforzamento del modello culturale dominante – quanto in senso positivo – ad esempio, orientando un *dataset* a sovra-rappresentare determinate minoranze, in una prospettiva di eguaglianza sostanziale[16].

Ai nostri fini è utile, infatti, considerare che i *bias* possono introdursi sia in sistemi di IA *rule-based*, dunque più semplici, con un funzionamento logico prevedibile ed esplicabile *ex ante*, nonché modificabile solo in fase di programmazione, ma anche, e maggiormente, nei sistemi di *machine learning,* in cui i risultati si basano su un modello probabilistico, sono dinamici e non prevedibili *ex ante,* ma eventualmente esplicabili *ex post* attraverso il ricorso al *reverse engineering*[17].

Per i soli algoritmi di *machine learning,* che, come si è visto, assumono decisioni stabilendo correlazioni probabilistiche tra variabili che non sempre corrispondono a relazioni puramente causali, è poi necessario fare attenzione alla c.d. *proxy discrimination*. I *proxy* sono variabili assunte come rilevanti ai fini del processo decisionale dall'algoritmo. Seppur in apparenza non sembrino fattori di discriminazione, ma appaiano neutri, sono, in realtà, statisticamente e storicamente associati a un dato fattore di discriminazione, producendo inevitabilmente a specchio un analogo esito discriminatorio[18]. Volendo fornire un esempio, in tal senso, si pensi a un sistema di IA che nel considerare i bisogni di assistenza non assuma come fattore rilevante il genere del paziente, ma lo stato di gravidanza, cioè una variabile intrinsecamente connessa al genere. Un altro esempio, forse ancor più lampante, sarebbe il caso di un sistema di IA che assumesse come *proxy* un'altezza minima del paziente pari a 1,70 m: anche in questo caso, poiché statisticamente una simile altezza si trova con maggiore frequenza all'interno della popolazione maschile, si verificherebbe una *proxy discrimination* a scapito del genere femminile[19].

È qui che, nel riflettere sulla tenuta del diritto antidiscriminatorio, che trova il suo fondamento all'art. 21 della Carta di Nizza e all'art. 10 TFUE[20], rispetto a tali tecnologie di IA, si può osservare che i modelli di correlazione illustrati nella *proxy discrimination* rimandano alla categoria dogmatica della discriminazione indiretta[21], la cui nozione di derivazione

---

[16] M. D'Amico, C. Nardocci, *Intelligenza artificiale e discriminazione di genere: rischi e possibili soluzioni*, in G. Cerrina Feroni, (a cura di), *AI Anthology. Profili giuridici, economici e sociali dell'intelligenza artificiale,* Bologna, 2022; C. Casonato, *Per una intelligenza artificiale costituzionalmente orientata*, in A. D'Aloia, *Intelligenza artificiale e diritto come regolare un mondo nuovo*, Milano, 2021, 142.

[17] G. Gaudio, *Le discriminazioni algoritmiche*, in *Lavoro, Diritti, Europa,* 1, 2024, 5; M. Peruzzi, *Il diritto antidiscriminatorio al test di intelligenza artificiale,* in *Labour Law*, 7, 1, 2021, 56.

[18] A. Lo Faro, *Algorithmic Decision Making e gestione dei rapporti di lavoro: cosa abbiano imparato dalle piattaforme*, in *Federalismi.it,* 25, 2022, 195; F. Zuiderveen Borgesius, *Discrimination, artificial intelligence, and algorithmic decision-makin, Council of Europe,* in *Arxiv,* 2025, 10 ss..

[19] Questo esempio rappresenta un'analogia rispetto alle modalità di selezione per l'accesso al concorso in polizia che vi erano in Grecia, come rese note dal Caso Kalliri dinanzi alla Corte di Giustizia dell'Unione Europea (Kalliri c. Ministro degli Interni della Grecia, sent. 18/10/2017). Per un approfondimento sul caso cfr. B. Profumo, *All'altezza delle Forze dell'ordine: accesso ai concorsi in polizia e requisito della statura minima comune per uomini e donne,* in *Dpce Online,* 34, 1, 2018.

[20] Per una ricostruzione della dimensione comunitaria del principio di eguaglianza e della non discriminazione cfr. *ex pluribus* A. D'Aloia, *Eguaglianza. Paradigmi e adattamenti di un principio "sconfinato",* in *Rivista AIC*, 4, 2021; O. Pollicino, *Di cosa parliamo quando parliamo di uguaglianza?* in P. Calvieri (a cura di), in *Divieto di discriminazioni e giurisprudenza costituzionale*, Torino, 2006, 439-470.

[21] La discriminazione indiretta si verifica quando un criterio apparentemente neutro va a produrre un effetto proporzionalmente più svantaggioso negli appartenenti a un dato gruppo protetto o vulnerabile. Sul punto si rimanda al volume M. Barbera A. Guariso (a cura di), *La tutela antidiscriminatoria. Fonti, strumenti, interpreti*, Torino,

giurisprudenziale è codificata ormai da decenni nel diritto comunitario e in particolare dalla direttiva n. 2000/43/CE, nonché della discriminazione per associazione di genesi giurisprudenziale[22].

Il principio di uguaglianza e di non discriminazione costituisce uno dei valori fondanti dell'ordinamento dell'Unione europea, che la Corte di giustizia ha progressivamente arricchito di contenuto, e che trova espressione, non solo negli artt. 20, 21 e 26 della Carta di Nizza, ma anche nell'art. 2 TUE, nonché nel già citato art. 10 TFUE, che impongono all'Unione il contrasto di ogni forma di discriminazione nella definizione e nell'attuazione delle proprie politiche[23]. I *bias* algoritmici, nella misura in cui producono effetti discriminatori su determinati gruppi, sembrano astrattamente riconducibili alle categorie già esistenti – quali le citate discriminazione indiretta di cui alla direttiva n. 2000/43/CE e la discriminazione per associazione – corredate dagli strumenti tradizionali come l'inversione dell'onere della prova, tutela antidiscriminatoria civile. Va detto che la natura algoritmica introduce, però, elementi strutturalmente nuovi che mettono in crisi tali strumenti. I meccanismi attraverso cui i sistemi di IA producono trattamenti differenziati presentano, infatti, caratteristiche peculiari, opacità, complessità statistica, utilizzo di variabili *proxy* e capacità di apprendimento dinamico. Nonostante ciò, gli effetti giuridicamente rilevanti paiono continuare a manifestarsi secondo schemi riconducibili, in larga misura, alle figure della discriminazione diretta, indiretta e per associazione elaborate dal diritto dell'Unione europea. La discriminazione algoritmica non sembra, dunque, imporre la necessità di delineare categorie completamente nuove, quanto piuttosto una rilettura evolutiva delle categorie tradizionali del diritto antidiscriminatorio. In particolare, l'opacità delle *black box* rende difficile provare il nesso causale tra l'algoritmo e l'esito discriminatorio, la *proxy discrimination* opera su variabili apparentemente neutre in modo sistematico e su scala massiva, e l'*accountability* risulta frammentata tra sviluppatore, *deployer* e utente finale. La vera novità sembra, dunque, riguardare non tanto l'emersione di nuove fattispecie, quanto le difficoltà probatorie e di accertamento derivanti dall'opacità dei processi decisionali automatizzati, che rendono necessario un adattamento degli strumenti di tutela e una maggiore attenzione ai profili di trasparenza, *explainability* e *accountability*[24]. Il legislatore europeo ha risposto con l'*AI Act*, quale strumento principe di costituzionalismo digitale[25], volto a garantire che lo sviluppo tecnologico avvenga nel rispetto della dignità umana, dell'eguaglianza e dei diritti fondamentali sanciti dalla Carta dei diritti fondamentali dell'Unione europea. La governance dell'IA diviene così uno dei terreni nei quali si manifesta la vocazione dell'Unione a configurarsi come una comunità di diritto fondata sulla centralità

2019; G. Cerrina Feroni, A. Bellizzi di San Lorenzo, *Discriminazione e manifestazione del pensiero*, in *Libera critica in libero stato*, Napoli, 2015.

[22] La Corte di Giustizia dell'Unione Europea (cfr. in particolare Corte giust., sentenza 31.7.2008, Coleman, causa C-303/06) nel fornire un'interpretazione estensiva del fattore discriminante ha elaborato la nozione di discriminazione per associazione, che si ha quando si subisce un trattamento sfavorevole dal momento in cui si viene associati con un'altra persona che possiede una «caratteristica protetta». Si veda sul punto M. Barbera A. Guariso, (a cura di), *La tutela antidiscriminatoria. Fonti, strumenti, interpreti*, cit., 57.

[23] Per uno studio sistematico del principio si rimanda a K. Lenaerts, *Exploring the Limits of the EU Charter of Fundamental Rights*, in *European Constitutional Law Review,* 2012; M. G. Militello, *Principio di uguaglianza e di non discriminazione tra Costituzione italiana e Carta dei diritti fondamentali dell'Unione Europea (artt. 3 Cost.; art. 20 e art. 21 Carta di Nizza), in Biblioteca '20 Maggio'*, 1, 2010; M. Barbera, *Il principio di eguaglianza nel sistema europeo "multilivello"*, in E. Paciotti (a cura di), *I diritti fondamentali in Europa,* Roma, 2011, 63-78.

[24] S. Wachter, B. Mittelstadt, C. Russell, *Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI,* in *Computer Law & Security Review,* 41, 2021.

[25] Sul costituzionalismo digitale si rimanda *ex multis* a G. De Gregorio, *Digital Constitutionalism in Europe,* Cambridge, 2022; O. Pollicino, G. De Gregorio, *The European Artificial Intelligence Act and Digital Constitutionalism*, 2024; G. Pitruzzella, *Il costituzionalismo digitale tra Stati Uniti e Europa,* in *Unione Europea e diritti,* 2, 2025.

della persona. Il Regolamento introduce, infatti, obblighi *ex ante* – si fa, a tal fine, riferimento all'art. 10 sulla qualità dei *dataset*, all'art. 13 sugli obblighi di trasparenza e informazione per i sistemi ad alto rischio e all'art. 26 sugli obblighi gravanti in capo al *deployer* anche per evitare eventuali discriminazioni – che segnano un paradigma nel solco di una prevenzione normativa. La disciplina dei sistemi ad alto rischio e le prescrizioni in materia di trasparenza, supervisione umana e qualità dei dati rappresentano, infatti, misure poste a prevenire esiti discriminatori e, pertanto, a tutela del principio di eguaglianza e del diritto a non subire discriminazioni sanciti dagli artt. 20 e 21 della Carta di Nizza.

3.* Bias *e discriminazione algoritmica in sanità. Si riportano di seguito casi di sistemi di IA affetti da bias applicati al settore sanitario, quali evidenziati dalle linee guida *Ethics and Governance of Artificial Intelligence for Health* pubblicate dall'OMS nel 2021 e il *Blueprint for an AI Bill of Rights* dell'Ufficio della Scienza e Tecnologia della Casa Bianca pubblicato nell'ottobre 2022, classificati da chi scrive sulla base della tripartizione per fasi di intrusione dei *bias*, come illustrata *supra*.

Per quanto attiene al *dataset* come fonte di *bias,* e in particolare al *bias* di esclusione, si evidenzia che la maggior parte delle tecnologie di IA utilizzate in dermatologia in Paesi Occidentali come Stati Uniti e Australia, ma anche in Europa, fino ad alcuni anni fa[26], risultava addestrata su immagini di persone con carnagione chiara, con l'inevitabile esito di marginalizzare la porzione di popolazione con tonalità di pelle più scura. In particolare, è emerso che alcuni algoritmi progettati per aiutare i dermatologi a diagnosticare lesioni cutanee cancerose, addestrati esclusivamente su dati relativi a popolazioni bianche, non producevano risultati accurati per pazienti di colore, aumentando i rischi per la loro salute. L'inadeguatezza dei dati riguardanti persone di colore, dovuta a diversi fattori, ha portato, dunque, a una sottorappresentazione delle minoranze etniche nel *dataset*[27].

Un caso di *bias* algoritmico legato alla configurazione del sistema di IA è, poi, quello individuato nel 2019 da un *team* di ricercatori statunitensi. Si trattava di un significativo *bias* razziale all'interno di un algoritmo ampiamente utilizzato nel sistema sanitario statunitense, che guidava le decisioni sanitarie basandosi sui costi, piuttosto che sulla gravità delle malattie, come indicatore del bisogno. Nonostante i pazienti di colore e quelli bianchi presentassero lo stesso livello di bisogno, di fatto, il sistema sanitario americano spendeva meno denaro per i pazienti di colore[28]. Per ritornare ai concetti presentati *supra* l'algoritmo individuava come *proxy* delle necessità dei pazienti i costi sostenuti, anziché le condizioni patologiche. Di conseguenza, l'algoritmo deduceva erroneamente che i pazienti bianchi fossero più malati rispetto ai pazienti di colore, seppure a parità di condizioni. I ricercatori hanno stimato che tale *bias* razziale abbia causato una riduzione del 50% nel numero di pazienti di colore che hanno ricevuto cure aggiuntive[29]. Sul piano giuridico, casi di questo tipo mostrano come l'impiego di variabili apparentemente neutre possa tradursi in trattamenti differenziati, imponendo di interrogarsi non solo sulla correttezza tecnica del modello, ma anche sulla sua compatibilità col diritto antidiscriminatorio. Soprattutto se i sistemi operano in contesti sanitari pubblici, casi analoghi possono ricadere, infatti, sotto l'ombrello protettivo di cui alla Direttiva 2000/43/CE. Un'altra conseguenza sul piano giuridico riguarda la responsabilità che, nel caso di specie, sembrerebbe in capo al *deployer* ai sensi dell'art. 26 *AI Act*. In generale,

[26] A. Lashbrook, *AI-driven dermatology could leave dark-skinned patients behind*, in *The Atlantic*, 16 Agosto 2018.
[27] OMS, *Ethics and governance of artificial intelligence for health: WHO guidance*, 2021, 55.
[28] *Ivi*, 2021, 54.
[29] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, *Dissecting racial bias in an algorithm used to manage the health of populations*, in *Science*, 366, 2019, 447- 453.

come appare lampante, simili distorsioni possono incidere sull'effettivo godimento del diritto alla salute e sul principio di uguaglianza nell'accesso alle cure, con inevitabili effetti di rilievo costituzionale.

I *bias* possono, inoltre, affliggere, sistemi predittivi col correlato rischio di *nudging* emotivo. Le tecnologie predittive possono, infatti, rappresentare una minaccia per l'autonomia individuale, valicando i limiti della tradizionale relazione medico-paziente. L'implementazione dell'IA, unita a tecniche come il *nudging*[30], *i.e.* la spinta gentile verso un dato obiettivo, ha il potenziale di trasformare un'iniziativa di promozione della salute in uno strumento in grado di influenzare significativamente le scelte individuali. Si pensi al caso in cui l'IA predittiva anticipi un elevato rischio di una particolare malattia per un individuo, sorge il dubbio – almeno a livello etico – se quella persona conservi il diritto di intraprendere comportamenti che potrebbero aumentare la probabilità dell'insorgenza della malattia prevista. Sebbene l'implementazione di algoritmi predittivi sia spesso animata da buone intenzioni, le problematiche connesse possono superare i benefici auspicati, come dimostrato da un algoritmo predittivo volto a valutare la probabilità di gravidanze adolescenziali. Dal 2017, tale algoritmo, sviluppato da Microsoft, è stato impiegato nella provincia di Salta, in Argentina, per valutare la probabilità di gravidanze adolescenziali, con l'intento dichiarato di mitigare il fenomeno[31]. Il sistema si basava su algoritmi che identificavano caratteristiche socio-economiche in individui potenzialmente a rischio di *drop out* scolastico e gravidanze precoci, per prevedere ciò considerava però *ex post* i dati soggettivi di persone nel momento in cui erano in stato di gravidanza o in cui avevano già abbandonato la scuola. L'algoritmo è stato fortemente criticato per la sua inadeguatezza, in quanto era addestrato su un *dataset* meramente femminile, forniva previsioni sensibili su adolescenti senza il consenso loro o dei genitori, compromettendo così la *privacy*[32] e in parte anche la loro autonomia decisionale[33]. Rivolgendosi a soggetti vulnerabili, il sistema rafforzava potenzialmente atteggiamenti e politiche discriminatorie[34], anziché prevenirle, configurandosi dunque come un *bias* del terzo tipo, introdotto dall'intervento umano e in particolare un *bias* di conferma[35].

4. IA emotiva e rischi per i diritti fondamentali. È ormai ampiamente riconosciuto che i sistemi di IA abbiano conquistato anche l'ambito, un tempo creduto intangibile, della sfera emotiva. Tra le nuove sfide di una società algoritmica[36], ontologicamente caratterizzata dalla

[30] Per l'impatto del *nudging* sulla tutela costituzionale della persona nella società digitale si rimanda a E. Di Carpegna Brivio, *Nudge e protezione costituzionale della persona nella società digitale,* in *Federalismi.it,* 3, 2025, 60-80; S. Cassese, *Exploring the Legitimacy of Nudging,* in A. Kemmerer C. Möllers, M. Steinbeis, G. Wagner, (a cura di), *Choice Architecture in Democracies. Exploring the Legitimacy of Nudging,* Baden-Baden, 2016, 241-245.

[31] P. Peña P, J. Varon, *Decolonising AI: A transfeminist approach to data and social justice*, in *Artificial intelligence: Human rights, social justice and development. Global Information Society Watch. Association for Progressive Communications*, 2019.

[32] Sul piano della tutela dei dati personali, il caso solleva profili rilevanti ai sensi dell'art. 22 GDPR, che riconosce all'interessato il diritto di non essere sottoposto a decisioni basate unicamente su trattamento automatizzato che producano effetti giuridici o incidano in modo significativo sulla sua persona. Nel caso di specie, essendo coinvolti soggetti minorenni e in assenza di consenso degli interessati o dei loro rappresentanti legali, la violazione appare tanto più grave, in quanto l'art. 8 GDPR subordina il trattamento dei dati dei minori al consenso del titolare della responsabilità genitoriale.

[33] Laboratorio de Inteligencia Artificial Aplicada, Universidad de Buenos Aires, *Sobre la predicción automática de embarazos adolescents*, 2018.

[34] J. Venturini, *Surveillance and social control: How technology reinforces structural inequality in Latin America*, in *London Privacy International,* 2019; J. Ortiz Freuler, C. Iglesias, *Algorithms and artificial intelligence in Latin America: A study of implementation by governments in Argentina and Uruguay*, in *Washington Dc: World Wide Web Foundation*, 2018.

[35] OMS, *Ethics and governance of artificial intelligence for health: WHO guidance*, 2021,53.

[36] A. Pajno *et al.*, *AI: profili giuridici. Intelligenza Artificiale: criticità emergenti e sfide per il giurista*, in *BioLaw Journal*, 3, 2019, 205 – 235, 206.

diffusione tecnologica, vi è la necessità di definire l'ambito dell'IA emotiva e di delineare le implicazioni etiche e giuridiche connesse a tale fenomeno, incluse le potenziali minacce alle libertà e ai diritti fondamentali. In particolare, uno dei problemi più studiati è il rischio di esiti discriminatori insiti nell'applicazione acritica di queste tecnologie. Tale questione assume una rilevanza particolare in settori caratterizzati da dati ultrasensibili come la sanità.

Partendo dal presupposto che, di fronte a un fenomeno tanto sfaccettato, lo sviluppo di una categoria dogmatica unitaria risulta particolarmente complesso, si tenterà di fornire una definizione ampia di IA emotiva. Nel *mare magnum* dei modelli riconducibili alla categoria concettuale dell'intelligenza artificiale, già qualificata come *umbrella term*[37], è parimenti arduo delineare confini definitori che ricomprendano il ventaglio fenomenologico in cui l'IA emotiva si manifesta. Se, in un contesto tanto ampio e interdisciplinare, è possibile aderire alla definizione in senso stretto di *emotional AI* come viene intesa in letteratura l'*affective computing*[38], ovvero elaborazione affettiva che rileva dati emotivi e mira a captare, riconoscere, interpretare e talvolta persino a manipolare le emozioni umane[39], occorre sottolineare che la conquista del campo emotivo da parte dei sistemi di IA non si è limitata a ciò. In senso lato, infatti, si può osservare un'interazione tra queste tecnologie emergenti e la sfera emotiva anche nei processi creativi legati ai cosiddetti sistemi generativi di IA, in particolare con riferimento alla risposta emotiva suscitata negli utenti dagli *output* e, soprattutto, dai prodotti della creatività nei campi dell'arte, un tempo esclusivo appannaggio della creatività umana[40].

Analogamente, rientrano nell'ambito dell'IA emotiva – intesa in senso lato – quelle tecnologie caratterizzate dalla capacità di instaurare una relazione con l'essere umano e incidere così sulla sua sfera emotiva. Un caso esemplare è rappresentato dai *social robots,* impiegati in contesti educativi, assistenziali e sanitari. Tali sistemi, per definizione, sono capaci di instaurare e mantenere relazioni con gli esseri umani e sviluppare competenze sociali[41]. Sulla base di tali premesse, ai fini del presente contributo, per IA emotiva si intenderà l'IA emotiva *lato sensu*.

Un discorso sull'IA emotiva non può prescindere da un ulteriore tema centrale. Si evidenzia,

---

[37] F. A. Raso *et al., Artificial Intelligence & Human Rights: Opportunities & Risks*, in *Berkman Klein Center Research Publication*, 6, 2018, 10; A. D'aloia, *Il diritto verso "il mondo nuovo". Le sfide dell'Intelligenza Artificiale*, in *BioLaw Journal*, 1, 2019, 3 – 31, 8.

[38] Il concetto di *affective computing* è stato inizialmente proposto da Rosalind Picard, cfr. R. W. Picard, *Affective Computing,* in *M.I.T Media Laboratory Perceptual Computing Section Technical Report Nr. 321*, 1995. La nozione, come proposta dalla studiosa, esorbita, in realtà, gli stretti confini a cui la dottrina tende a circoscriverla e, secondo la sua definizione primigenia, può essere interpretata come ricomprendente tutti i casi in cui vi è un'interazione fra macchina ed emotività umana. Si consideri, inoltre, che nel 2009 il MIT di Boston fondò *Affectiva*, società che sviluppa sistemi di IA specializzati nella rilevazione di emozioni partendo dall'analisi dei dati biometrici come movimenti del corpo, del volto e tono della voce. Per una ricostruzione del concetto di *affective computing* si veda S.B. Daily *et al.*, *Affective Computing: Historical Foundations, Current Applications, and Future Trends*, in M. Jeon (a cura di), *Emotions and affect in human factors and human-computer interaction*, Londra, 2017, 213 ss.; Y. Wang *et al.*, *A systematic review on affective computing: emotion models, databases, and recent advances,* in *Information Fusion,* 83-84, 2022, 19-52.

[39] Y. Liu *et. al.*, *Affective Computing for Healthcare: Recent Trends, Applications, Challenges, and Beyond*, in *Human-Computer Interaction*, 2024; T. Gremsl, E. Höld, *Emotional AI: Legal and Ethical Challenges,* in *Information Polity*, 2, 2022, 163 – 174, 164.

[40] Si cita in proposito, a titolo esemplificativo, lo studio dell'Università di Pittsburgh da cui è emerso che lettori non esperti a cui venivano sottoposti sia testi poetici generati dall'IA, sia testi di poeti classici come Shakespeare e Dickinson tendevano a considerare più belle le poesie generate dall'IA, anziché quelle di autori classici e ad ascrivere la paternità di poesia generate dall'IA ai grandi poeti. Cfr. B. Porter, E. Machery, *AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably,* in *Scientific Reports,* 14, 2024.

[41] Per un'analisi sulla questione dell'interazione relazionale tra tecnologia e esseri umani e sulle implicazioni emotive, in settori come quello della salute e dell'assistenza alle fasce vulnerabili, si rimanda ad A. D'Aloia, *Intelligenza artificiale, società algoritmica, dimensione giuridica. Lavori in corso,* cit., *passim.*

infatti, come, partendo da una concezione inevitabilmente antropocentrica di intelligenza, la dottrina prevalente abbia ritenuto che l'intelligenza umana si distinguesse da quella artificiale in virtù della presenza di un *quid pluris,* identificabile nella coscienza o nell'esperienza[42]. Alla luce del potenziamento esponenziale, almeno in termini quantitativi, della capacità dei sistemi di IA di replicare materiale emotivo, l'esito inevitabile è l'ipotesi di un possibile superamento futuro della suddetta dicotomia. Tale fatto, in linea con le teorie già kelseniane (cioè la persona come *fictio*, fascio di norme[43]), viene utilizzato come argomento, seppure in via ancora meramente speculativa, a favore di un possibile riconoscimento della soggettività giuridica di alcuni sistemi di IA[44].

Altre rilevanti questioni giuridiche che caratterizzano il dibattito sui sistemi di IA emotiva sono le seguenti. In primo luogo, si consideri il dato fattuale per cui l'interpretazione o la riproduzione delle emozioni richiede un passaggio preliminare, ossia il riconoscimento da parte del sistema delle emozioni dell'interlocutore, il quale avviene sulla base di indicatori biometrici, comportamentali e fisiologici[45]. Ciò solleva possibili problematiche connesse al diritto alla privacy, alla libertà di autodeterminazione e di espressione delle proprie emozioni. I dati emotivi si configurano come dati ultrasensibili, rientranti sotto l'ombrello protettivo dell'art. 21 Cost.[46], in quanto materializzazione della libertà di espressione e manifestazione intima della personalità. La suddetta libertà risulta potenzialmente compressa dall'eventuale consapevolezza di essere analizzati sulla base dell'espressione delle proprie emozioni, suggerendo così la necessità di delineare una sorta di *habeas mentem*.

La questione centrale, che rileva maggiormente ai nostri fini e che verrà analizzata *infra*, riguarda, però, il rischio di esiti discriminatori derivanti dall'uso acritico di tali sistemi, in violazione del principio di uguaglianza. Le conseguenze discriminatorie potenziali derivano dalla capacità di queste tecnologie di rilevare o riconoscere accuratamente solo le emozioni appartenenti alla cultura dominante o a coloro che hanno programmato il modello, discriminando così le classi sottorappresentate (c.d. imperialismo emotivo)[47]. Le modalità di espressione delle emozioni variano sensibilmente da cultura a cultura e l'universalità dei modi di esprimere le emozioni è stata a più riprese confutata, in quanto eccessivamente riduzionistica. Barrett, professoressa di psicologia alla Northeastern University, ha evidenziato, ad esempio, che l'espressione comunemente classificata come paura – secondo la classificazione di studi psicologici occidentali – è invece interpretata come rabbia o minaccia dai Maori della Nuova Zelanda e dagli abitanti delle Isole Trobriand della Papua

[42] Sul tema *ex multis* L. B. Solum, *Legal Personhood for Artificial Intelligences*, in *North Carolina Law Review*, 70, 4, 1992, 1264-1266.

[43] «*Thus the concept of legal subject or legal person is simply a personifying expression for the unity of a bundle of legal obligations and legal rights, that is, the unity of a complex of norms*» H. Kelsen, *Introduction to the Problems of Legal Theory (the first edition of Reine Rechtslehre*), Oxford, 1992, 47.

[44] A. Santosuosso, M. Tomasi, *Diritto, scienza, nuove tecnologie*, Milano, 2021, 185; A. D'Aloia *Intelligenza artificiale e diritto come regolare un mondo nuovo*, cit., 60.

[45] S. Yaqoob *et al., AIoT for human emotion recognition: Potentials, challenges, and healthcare applications,* in *Computer Science Review,* 60, 2026.

[46] Così, ad esempio, Sergio Fois «il pensiero considerato nell'articolo 21 non deve essere inteso in senso rigido e stretto: è garantita anche la manifestazione di suggestioni, emozioni, sensazioni, nonché qualsiasi attività volta a provocare corrispondenti stati d'animo» cfr. S. Fois, A. Vignudelli (a cura di), *Principi fondamentali, diritti e doveri dei cittadini nella Costituzione italiana, Articoli* 1-54, Rimini, 1991, 55; P. Tanzarella, *Definizione e ridefinizione della libertà d'espressione nei settant'anni di vigenza della Costituzione italiana,* in *Consulta Online, Liber Amicorum per Pasquale Costanzo,* 14 maggio 2020.

[47] S. Latif *et al., AI-Based Emotion Recognition: Promise, Peril, and Prescriptions for Prosocial Path,* in *Human – Computer Interaction*, 14, 2022, 5.

Nuova Guinea[48]. Alla luce della riconosciuta infondatezza scientifica e del superamento della *Basic Emotion Theory* (BET), sviluppata dallo psicologo Ekman negli anni '60, che sosteneva la possibilità di dedurre universalmente gli stati emotivi delle persone osservando espressioni e micro-espressioni facciali, si ritiene che non vi sia una corrispondenza assoluta tra espressione ed emozione. I modelli di IA per il riconoscimento emotivo rischiano di fondarsi su teorie né universalmente riconosciute né valide in modo assoluto, che da decenni sono oggetto di critiche, anche per il rischio di un appiattimento dei sentimenti. Di conseguenza, i dati assunti come indicatori non dovrebbero essere considerati universalizzabili: una data espressione non può essere univocamente associata a una specifica emozione, poiché le modalità espressive variano non solo tra gruppi etnici o socio-culturali diversi, ma anche tra individui dello stesso gruppo e persino nei diversi momenti della vita o della giornata di un singolo individuo[49]. L'idea di leggere e interpretare le emozioni in forma digitale ha portato a studi sempre più mirati per identificare una sorta di formula matematica dell'emozione, dimenticando che la natura inevitabilmente individuale dei dati emotivi rende i sistemi di IA emotiva particolarmente inclini al rischio di *bias*. Ricerche recenti indicano, infatti, che le tecnologie di analisi emotiva tendono ad attribuire emozioni negative con maggior frequenza a individui appartenenti a specifiche etnie e che il livello attuale di accuratezza dell'IA non consente di cogliere le sfumature culturali nell'esprimere e interpretare le emozioni, rendendo difficile giungere a valutazioni precise[50]. Questo accade in particolare quando i *dataset* utilizzati per addestrare l'IA emotiva sono condizionati demograficamente. In tali casi, l'affidabilità dell'IA diminuisce nel momento in cui essa viene estesa a una popolazione più ampia, poiché molti sforzi di raccolta dati si sono storicamente concentrati su specifici set di dati demografici, ossia principalmente su individui che si sono offerti volontari per le sperimentazioni o che potevano permettersi certe tecnologie. Partendo dalla convinzione che casi simili comportano un'interferenza nella sfera intangibile dell'integrità individuale, ulteriori aspetti rilevanti – già menzionati anche nel documento del Consiglio d'Europa sull'AI, Raccomandazione 2102/2017[51] – riguardano il rispetto della dignità umana e il suo corollario dell'autodeterminazione, frustrato dall'uso manipolativo dei dati per indirizzare i comportamenti degli utenti. La dignità umana si configura come concetto relazionale[52], da intendersi, ai nostri fini, sia nel senso giusnaturalistico adottato dall'art. 1 della Dichiarazione Universale dei Diritti dell'Uomo, come substrato inviolabile dei diritti umani[53], sia nel senso kantiano[54] dell'irriducibilità dell'uomo a una sua mera parte fenomenica[55], *rectius*, in questi casi, alle sue emozioni o ai dati biometrici che si assume siano rivelatori di queste. Si evidenziano, inoltre, potenziali preoccupanti analogie tra i sistemi di riconoscimento emotivo e discipline pseudoscientifiche, come la frenologia e la fisiognomica[56], che rimandano

[48] L. F. Barrett *et al., Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements*, in *Psychological Science in the Public Interest,* 20, 2019, 1-68; M. Martorana, R. Savella, *IA e riconoscimento delle emozioni: rischi e possibili vantaggi,* in *Agenda Digitale*, 21 Settembre 2023.

[49] M. Purdy *et al., The Risks of Using AI to Interpret Human Emotions,* in *Harvard Business Review Home*, 18 Novembre 2019.

[50] *Ibidem.*

[51] Recommendation 2102(2017), *Technological convergence, artificial intelligence and human rights*, Council of Europe, 2017.

[52] F. Scamardella, *La dimensione relazionale come fondamento della dignità umana,* in *Rivista di Filosofia del Diritto*, 2, 2013, 305-320.

[53] Per un approfondimento si veda *ex pluribus* A. Apostoli, *La dignità sociale come orizzonte della uguaglianza nell'ordinamento costituzionale,* in *Costituzionalismo.it*, 3, 2019, 2-32.

[54] I. Kant, *Fondazione della metafisica dei costumi,* Konisberg, 1785, trad. it. 55.

[55] G. Azzoni, *Nomofanie. Esercizi di filosofia del diritto,* Torino, 2021, 241.

[56] D. Huyskes, *Riprendiamoci le emozioni dal controllo dell'intelligenza artificiale*, in *La Repubblica,* 1 giugno 2021.

all'antropologia criminale e al positivismo lombrosiano.
Volendo indagare come il legislatore europeo ha affrontato il fenomeno dell'IA emotiva – e, nello specifico, dei sistemi di riconoscimento delle emozioni, di cui fornisce una definizione[57] all'art. 3, n. 39[58] – v'è da dire che il regolamento 2024/1689 si fonda sull'assunto che determinati sistemi di IA, come i dispositivi medici, i sistemi che influenzano la salute e i sistemi di IA emotiva – non vietati in ambito sanitario – siano generalmente ad alto rischio, a causa del potenziale impatto negativo sulla sicurezza e sui diritti fondamentali[59]. L'*Ai Act*, nell'adottare il noto approccio basato sul rischio, identifica un rischio intrinseco nel riconoscimento delle emozioni e, in particolare, un rischio inaccettabile con conseguenti divieti e limiti stringenti nell'uso di sistemi di riconoscimento emotivo nei luoghi di lavoro[60] e nell'istruzione, escludendo espressamente da tale divieto, però, i sistemi legati a motivi medici o di sicurezza[61]. Individua, inoltre, un rischio specifico per la trasparenza e la necessità di essere informati quando si utilizzano sistemi di IA, come *chatbot*, integrati dal riconoscimento emotivo[62]. Anche se il divieto di immissione sul mercato, messa in servizio o uso dei sistemi di IA emotiva non dovrebbe riguardare i modelli destinati strettamente a scopi medici o di sicurezza, come i sistemi a uso terapeutico, il considerando 44 dell'*Ai Act* esprime preoccupazioni rilevanti circa l'affidabilità scientifica di tali modelli. Tra i principali punti critici di tali sistemi figurano l'affidabilità limitata, la scarsa specificità e la ridotta generalizzabilità. Pertanto, i sistemi di IA che identificano o inferiscono emozioni o intenzioni di persone fisiche sulla base dei loro dati biometrici possono dar luogo a esiti discriminatori e risultare intrusivi rispetto ai diritti e alle libertà degli interessati[63]. Non è un caso che siano vietati nel contesto lavorativo o educativo, proprio perché potrebbero determinare trattamenti pregiudizievoli o sfavorevoli verso determinate persone fisiche o interi gruppi[64].

***5. IA emotiva applicata al settore sanitario e* bias.** Numerose iniziative nel campo della salute digitale stanno integrando tecnologie basate sull'IA emotiva, e si prevede che la loro applicazione sia destinata ad aumentare progressivamente. Sebbene il numero di articoli

[57] È degno di nota che il testo definitivo del Regolamento, diversamente dalla bozza precedente alla fase di trilogo, preveda all'art. 3 – tra 68 definizioni – una definizione esplicita di sistema di riconoscimento delle emozioni, enumerando, nel considerando 18, anche una serie di emozioni ed escludendo, a tali fini, stati fisici oggettivi come la fatica e il dolore. A integrazione *Cfr.* Regolamento (UE) 2024/1689, Considerando 18.
[58] Regolamento (UE) 2024/1689, Art. 3 n. 39 «per *sistema di riconoscimento delle emozioni,* si deve intendere *un sistema di IA finalizzato all'identificazione o all'inferenza di emozioni o intenzioni di persone fisiche sulla base dei loro dati biometrici*».
[59] Cfr. Regolamento (UE) 2024/1689, Art. 6, paragrafo 2 e allegato III.
[60] A proposito degli effetti lesivi di sistemi di IA emotiva sul luogo di lavoro, si riporta un caso paradigmatico. Nel 2021, l'azienda cinese *Canon Information Technology* ha installato telecamere intelligenti per riconoscere il sorriso dei dipendenti al fine di determinare se ammetterli o no sul luogo di lavoro. Per un approfondimento sul caso si veda D. Ruggiu, *L'emozione, nuovo territorio di conquista dell'intelligenza artificiale: applicazioni e rischi*, in *Agenda Digitale*, 27 Luglio 2021.
[61] Cfr. Regolamento (UE) 2024/1689, art. 5, lett. f).
[62] Cfr. Regolamento (UE) 2024/1689, art. 50, c. 3.
[63] Alla stregua di quanto finora detto, si evidenzia che il Considerando 44 specifica che vi sono serie perplessità sulla natura intrusiva e sulla base scientifica dei sistemi di IA che mirano a identificare o inferire emozioni, in particolare poiché le espressioni emotive variano considerevolmente tra culture, situazioni e persino all'interno dello stesso individuo. Cfr. Regolamento (UE) 2024/1689, Considerando 44.
[64] Come emerge dalle linee guida sulle pratiche vietate elaborate dalla Commissione Europea, il luogo lavorativo è da intendersi in un'accezione ampia, inoltre, la limitazione del divieto ai casi di contesti lavorativi ed educativi va ricondotta all'asimmetria di poteri che in tali contesti si può verificare. Cfr. Communication from the Commission, *Approval of the content of the draft Communication from the Commission - Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act)*, C(2025) 884, 4 Febbraio 2025.

scientifici pubblicati in tema di *affective computing* sia in costante aumento ogni anno – tendenza, questa, osservata anche in campi come la salute digitale – quando si cercano articoli che combinano *affective computing* o IA emotiva e salute, il numero di pubblicazioni rilevanti disponibili diminuisce significativamente. Da ciò emerge che vi è ancora una mancanza di analisi sistematica dei recenti progressi dell'IA emotiva nell'assistenza sanitaria[65]. Sebbene non esista una definizione formale di salute digitale affettiva (o emozionale), si può dire che tale concetto emergente mette in relazione la salute digitale e l'intelligenza artificiale al fine di captare bisogni e risposte efficaci, anche sulla base delle emozioni dei pazienti[66].

Negli ambienti sanitari, la raccolta e l'annotazione dei dati emotivi sono spesso influenzate da *bias* intrinseci dovuti, sia alla differenza delle popolazioni nell'esprimere le emozioni, sia alla soggettività degli annotatori umani. Le diverse preferenze e preoccupazioni dei singoli operatori sanitari durante la raccolta e il *labeling* dei dati contribuiscono a determinare significativi *bias*, rendendo ancora più lontano l'obiettivo di un addestramento di modelli di IA equi. V'è da dire che, per affrontare queste sfide e ottenere una comprensione imparziale del dato emotivo, i ricercatori stanno esplorando approcci innovativi tra cui l'uso di algoritmi di apprendimento non supervisionato e auto-supervisionato, che possono apprendere rappresentazioni emotive da dati non etichettati, riducendo così la soggettività legata alla fase di *labeling*[67].

Un esempio di sistema di IA emotiva *stricto sensu*, utilizzato nel contesto della salute psicofisica e accessibile alla generalità dei consociati, è il sistema *TWILL* (prima *Happify*), dotato di una piattaforma di guarigione intelligente che impiega l'IA per comprendere le esigenze di salute di un individuo e raccomandare azioni appropriate tramite un *chatbot* sanitario addestrato a fornire cure e supporto personalizzati ed empatici. In modo non dissimile agisce, poi, *LUCID*, un sistema di raccomandazione che suggerisce musica basata sullo stato psicofisico dell'utente, incorporando dati biometrici e auto-valutati come *input* per classificare gli stati emotivi dell'utente e fornirgli un'esperienza personalizzata[68].

Per quanto riguarda l'intelligenza artificiale emotiva *lato sensu*, si noti che i *bias* possono essere osservati anche in sistemi di IA generativa e nei robot sociali. Le tecnologie di IA generativa soffrono di potenziali *bias* dovuti, ad esempio, al *dataset* di addestramento degli algoritmi. È agevole immaginare che un sistema di IA generativa potrebbe fornire risultati iniqui, ad esempio, perché influenzato da *bias* linguistici, a causa di quantità variabili di dati disponibili nelle diverse lingue[69]. Una prima applicazione di *generative AI* nel settore sanitario è, a titolo esemplificativo, *Syntegra Medical Mind*[70], un sistema che crea record di pazienti sintetici realistici, provenienti da dati sanitari reali, da cui gli operatori sanitari possono attingere, proteggendo, però, la privacy degli individui. Nel fare ciò, affronta anche il problema dei dati *biased* e promuove l'eguaglianza algoritmica nonché piani di trattamento equi[71]. È stato poi dimostrato[72] che anche sistemi di IA generativa di immagini con finalità generali, come *DALL-E 2,* possono essere applicati fruttuosamente in radiologia, in quanto

[65] Y. Liu et. al., *Affective Computing for Healthcare: Recent Trends, Applications, Challenges, and Beyond,* cit.

[66] J. Apablaza, S. Cano, *Affective Computing from Digital Health: A literature review,* in *Workshop on ICTs for improving Patients Rehabilitation Research Techniques,* 2022.

[67] Y. Liu et. al., *Affective Computing for Healthcare: Recent Trends, Applications, Challenges, and Beyond,* cit.

[68] A. Labbé, *Emotion AI: Why It's The Future Of Digital Health,* in *Forbes*, 23 Novembre 2022.

[69] L. Bertoldini, *L'IA generativa è addestrata principalmente in inglese, e questo è un problema,* in *Facta*, 15 Settembre 2025.

[70] Per i progressi in campo medico grazie a dati sintetizzati col sistema cfr. G. Muniz-Terrera, *Virtual cohorts and synthetic data in dementia: An illustration of their potential to advance research,* in *Front. Artif. Intell,* 4, 2021.

[71] P. Zhang, *Generative AI in Medicine and Healthcare: Promises, Opportunities and Challenges*, in *Future Internet*, 15(9), 286, 2023, 7.

[72] L. C. Adams *et al.*, *What Does DALL-E 2 Know About Radiology?,* in *J. Med. Internet Res,* 25, 2023.

potenzialmente in grado di creare immagini a raggi X con stile e proporzioni anatomiche simili a quelle autentiche. In particolare, infatti, è stata appurata la sua capacità di generare immagini mediche a partire da *prompt* brevi e di ricostruire immagini radiologiche con aree mancanti. È noto, però, che *DALL-E* soffre di pregiudizi[73] nei suoi dati di addestramento e che, nonostante i *bias* si stiano progressivamente mitigando, in numerosi casi genera immagini che non riflettono accuratamente le diversità della popolazione mondiale[74]. Nel novero dell'IA generativa rientrano anche *chatbot* usati per il supporto psicologico come il promettente *chatbot* terapeutico *Therabot*[75], che interagisce direttamente con la sfera emotiva dei pazienti. Si riportano, in proposito, i dati del *trial* del 2025 della *Dartmouth Geisel School of Medicine*, che ha verificato un miglioramento fino al 51% delle condizioni di pazienti con depressione maggiore, disturbo d'ansia generalizzato e rischio clinico di disturbi alimentari[76]. Esempi virtuosi come il suddetto potrebbero compensare la carenza di professionisti della salute mentale e il loro sovraccarico di lavoro[77]. Da questo punto di vista è, poi, necessario fare riferimento all'*AI Act* che oltre a stabilire, come si è chiarito, l'obbligo della trasparenza per cui l'interlocutore deve essere a conoscenza che sta interagendo con un *chatbot*, *ex* art. 50 c. 1, introduce all'art. 5, c. 1, lett. a) e lett. b), in proposito, determinati divieti in relazione a pratiche di IA manipolative e che sfruttino la vulnerabilità delle persone[78]. L'*AI Act*, è bene ribadirlo, dunque, non vieta i *chatbot* terapeutici ed emotivi ma, come da approccio costitutivo dello stesso regolamento, li classifica in base al rischio, prevedendo che un *chatbot* emotivo possa qualificarsi come «sistema ad alto rischio» se applicato in uno dei contesti specificamente elencati nell'Allegato III, come quello sanitario in senso stretto, dunque, qualora svolga funzioni che incidono sulla salute mentale o interagisca con utenti vulnerabili. Più precisamente, però, le pratiche manipolative o che sfruttino la vulnerabilità degli utenti risultano *ex se* vietate ai sensi dell'art. 5, c. 1, lett. a) e b), indipendentemente dalla classificazione dei sistemi.

[73] Per una panoramica sui *bias* di DALL-E 2 si rimanda a N. Lubin, Lost in *DALL-E 2 AI Creativity is biased,* in *Medium,* 13 Agosto 2022; sui *bias* linguistici si veda, ad esempio, Y. Jun, *Lost in DALL-E 3 Translation,* in *Ux Magazine,* 7 Dicembre 2023.

[74] P. Zhang, *Generative AI in Medicine and Healthcare: Promises, Opportunities and Challenges*, cit., 6.

[75] M. V. Heinz, *Randomized Trial of a Generative AI Chatbot for Mental Health Treatment,* in *NEJM AI,* 2, 4, 2025.

[76] C. Maurizio, *Chatbot terapeutici: rivoluzione o placebo per la salute mentale?,* cit.

[77] In proposito va, però, detto che recenti studi sostengono che, allo stato attuale, soltanto un terapeuta umano avrebbe la capacità di cogliere davvero l'esperienza soggettiva della persona e il suo contesto emotivo e sociale e di instaurare una relazione terapeutica autentica. Si tratta di competenze relazionali e interpretative che i *chatbot* progettati per scopi terapeutici non sono ancora in grado di eguagliare, in particolare per quanto concerne la capacità di raggiungere la profondità emotiva necessaria per sviluppare un rapporto di fiducia alla base dell'alleanza terapeutica. In questo senso, cfr. Z. Zhang, J. Wang, *Can AI replace psychotherapists? Exploring the future of mental health care,* in *Frontiers in Psychiatry,* 15, 2024; L. Spytska, *The use of artificial intelligence in psychotherapy: development of intelligent therapeutic systems,* in *BMC Psychology,* 13, 175, 2025.

[78] Così l'art. 5, c. 1, Regolamento (UE) 2024/1689.

«Sono vietate le pratiche di IA seguenti:

a) l'immissione sul mercato, la messa in servizio o l'uso di un sistema di IA che utilizza tecniche subliminali che agiscono senza che una persona ne sia consapevole o tecniche volutamente manipolative o ingannevoli aventi lo scopo o l'effetto di distorcere materialmente il comportamento di una persona o di un gruppo di persone, pregiudicando in modo considerevole la loro capacità di prendere una decisione informata, inducendole pertanto a prendere una decisione che non avrebbero altrimenti preso, in un modo che provochi o possa ragionevolmente provocare a tale persona, a un'altra persona o a un gruppo di persone un danno significativo;

b) l'immissione sul mercato, la messa in servizio o l'uso di un sistema di IA che sfrutta le vulnerabilità di una persona fisica o di uno specifico gruppo di persone, dovute all'età, alla disabilità o a una specifica situazione sociale o economica, con l'obiettivo o l'effetto di distorcere materialmente il comportamento di tale persona o di una persona che appartiene a tale gruppo in un modo che provochi o possa ragionevolmente provocare a tale persona o a un'altra persona un danno significativo».

Un altro esempio è rappresentato da *chatbot* come *ChatGPT* e *Bard di Google*. Data la crescente dipendenza dei medici dai sistemi di intelligenza artificiale per compiti di routine, uno studio recente della Stanford School of Medicine solleva preoccupazioni riguardo a *chatbot* ampiamente utilizzati dai medici che perpetuano idee scientifiche infondate e di matrice razzista. Questi *chatbot,* addestrati su vasti testi internet, hanno mostrato concezioni e informazioni false sui pazienti di colore in risposta alle richieste dei ricercatori, potenzialmente acuendo problemi già radicati di razzismo medico[79].

Anche per quanto riguarda i *social robots*, si evidenzia che sussistono preoccupazioni sulla riproduzione di *bias* tipicamente umani negli algoritmi in questi innestati. Si pensi, in particolare, al caso di *carebots*, i.e. robot applicati all'assistenza sanitaria. Gli algoritmi implementati in tali robot spesso si basano su dati forniti da esperti umani, replicando inevitabilmente i loro *bias*[80], o mostrano una tendenza a categorizzare i pazienti in base a somiglianze con altri[81]. È necessario sottolineare che un *bias* potrebbe influenzare anche i pazienti ed essere diretto verso i robot. Per quanto riguarda i *bias* cognitivi e sociali dei pazienti nei confronti dei robot utilizzati nell'assistenza sanitaria, le ricerche indicano un aumento della fiducia quando i partecipanti interagiscono con terapisti robotici rispetto ai terapisti umani[82].

Il recente studio neozelandese[83] *Robots and racism* suggerisce persino che tendiamo a percepire i robot umanoidi come appartenenti a un'etnia e quindi applichiamo agli stessi stereotipi razziali[84]. Altrettanto attuali discussioni in tema di roboetica[85] arrivano a evidenziare discriminazioni nei confronti dei robot basate sul loro *design* fisico. Anche se *ictu oculi* potrebbe sembrarci che una discriminazione indirizzata a un robot sia un'eventualità che non si sostanzia in un danno sociale, si evidenzia che molteplici attori nel contesto sociale potrebbero subire conseguenze negative. Un possibile danno è il c.d. *witness harm*, ovvero il danno da testimone o danno riflesso, che ricade su una terza parte che assista alla discriminazione in parola, sperimentandone le conseguenze negative. Qualora questo tipo di discriminazione fosse tollerata, inoltre, le discriminazioni già radicate nel tessuto sociale potrebbero essere ulteriormente esacerbate[86]. Un altro aspetto cruciale riguarda le fasce di popolazione più vulnerabili, come bambini, anziani o persone con disabilità. Tali categorie potrebbero risultare particolarmente inclini a riporre un'eccessiva fiducia nelle interazioni sociali coi robot. Gravi problemi sorgono quando la fiducia nei confronti di sistemi di IA è calibrata in modo scorretto e, se combinata con algoritmi viziati da *bias*, il rischio di danno aumenta notevolmente[87]. Sul punto, è fondamentale sottolineare, infatti, che sistemi capaci di esibire comportamenti realistici, rispondere a gesti sociali e utilizzare espressioni facciali

[79] J. A. Omiye *et. al., Large language models propagate race-based medicine*, in *Digital Medicine*, 6, 195, 2023.

[80] Si consideri, a titolo esemplificativo, il caso di un robot riabilitativo che nell'interagire con una giovane paziente escluda dalla terapia l'uso di un videogioco riabilitativo, in quanto considerato più idoneo a pazienti maschi, basandosi, dunque, su un *dataset* di addestramento non accurato, affetto da un *bias* di genere di lunga data.

[81] A. Howard, J. Borenstein, *Trust and bias in Robots*, in *American Scientist*, 107, 2, 2019, 86-89.

[82] *Ibidem*.

[83] Lo studio *Robots and racism* condotto dallo Human Interface Technology Laboratory in Nuova Zelanda (HIT Lab NZ) e pubblicato dall'Università di Canterbury.

[84] C. Bartneck et. al., *Robots and racism*, in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018.

[85] Termine coniato dall'ingegnere robotico Gianmarco Verruggio. Sul tema si rimanda a *Roboethics: Ethics Applied to Robotics,* in *IEEE Robotics & Automation Magazine,* 18, 2011.

[86] J. K. Barfield, *Discrimination against robots: Discussing the ethics of social interactions and who is harmed*, in *Paladyn, Journal of Behavioral Robotics*, 14, 2023.

[87] P. Robinette et. al., *Overtrust of robots in emergency evacuation scenarios*, in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, 2016, 101–108.

per comunicare emozioni in un modo percepito come intrinsecamente umano, prendono di mira le risposte biologiche involontarie dell'essere umano e ne influenzano notevolmente la percezione[88]. A tal proposito, il costituzionalista statunitense Balkin fa riferimento al cosiddetto *substitution effect* (effetto sostitutivo), che riguarda la trasformazione delle relazioni sociali man mano che i robot, gli agenti di IA e gli algoritmi sostituiscono progressivamente l'essere umano in diverse funzioni, diventando così delle *special-purpose people*[89]. La funzione sostitutiva assunta da tali sistemi genera una tendenza umana a proiettare valori ed emozioni su di essi, soprattutto nei casi in cui questi sistemi possiedano elevate capacità di replicazione delle caratteristiche sensoriali e motorie proprie della natura umana. Balkin chiarisce, tuttavia, che questo effetto sostitutivo è parziale, incoerente e instabile e opportunistico, portando i robot sociali a essere percepiti alternativamente come persone, animali o oggetti, a seconda del contesto[90].

6. Mitigazione dei* bias *e IA come strumento di democratizzazione dei processi decisionali. Per mitigare i *bias* nei sistemi di IA applicati all'ambito sanitario, vengono ad oggi prospettate fondamentali misure e strategie quali in particolare: la diversificazione dei *dataset,* la revisione e il monitoraggio regolare e costante degli algoritmi e un'adeguata formazione dei pazienti e degli operatori sanitari. Uno dei metodi più efficaci per ridurre i *bias* è l'utilizzo di *dataset* diversificati e rappresentativi nella fase di sviluppo e addestramento dell'IA. Ciò implica la raccolta di dati da una vasta gamma di fonti per riflettere le caratteristiche demografiche, i bisogni sanitari e le eventuali disuguaglianze della popolazione presa a *target* tutta. Lo sviluppo di un *dataset* diversificato garantisce che i sistemi di IA possano rispondere a esigenze diverse dei pazienti, promuovendo maggiore sicurezza e fiducia nelle soluzioni sanitarie *AI based*. Anche l'art. 10 dell'*AI Act*[91] stabilisce, infatti, l'obbligo di impiegare *dataset* di addestramento affidabili, completi e rappresentativi, imponendo anche l'uso di strumenti dedicati all'analisi e alla validazione per individuare e correggere eventuali *bias* idonei a incidere sulla salute e sulla sicurezza delle persone, o a comportare discriminazioni vietate dal diritto dell'Unione[92]. Anche il riequilibrio dei dati, ottenuto tramite tecniche di *oversampling* e *undersampling*[93], può contribuire a ridurre le distorsioni e a rendere più equi i modelli di *machine learning*[94]. Una strategia è proprio la progettazione etica di *database* variegati sin dalle prime fasi dello sviluppo dei sistemi. In questo contesto, assumono particolare rilevanza i dati sintetici[95], i.e. dati sintetizzati artificialmente da algoritmi che replicano le caratteristiche statistiche dei dati reali, senza includere dettagli privati. L'impiego di tali dati, oltre a essere una soluzione più economica rispetto alle raccolte di dati effettive,

[88] Un esperimento significativo condotto nel 2013 da ricercatori del MIT Media Lab del Massachusetts ha dimostrato che, dopo aver trascorso un paio d'ore in una stanza con dinosauri robotici simili a cuccioli che reagivano al comportamento umano con movimenti realistici, i partecipanti si rifiutavano di colpirli violentemente, nonostante le richieste degli sperimentatori.

[89] J. M. Balkin, *The Three Laws of Robotics in the Age of Big Data*, in *Ohio State Law Journal,* 78, 1217, 2017, 14.

[90] J. M. Balkin, *The Path of Robotics Law,* in *California Law Review Circuit,* 6, 2015, 56; A. D'Aloia, *Intelligenza artificiale, società algoritmica, dimensione giuridica. Lavori in corso,* in *Quaderni Costituzionali*, 3, 2022, 675.

[91] Regolamento (UE) 2024/1689, art. 10.

[92] Cfr. Nello specifico Regolamento (UE) 2024/1689, art. 10, c. 2 lett. f.

[93] L'*oversampling* è la tecnica che permette di gestire set di dati sbilanciati, aumentando la rappresentanza della classe minoritaria, rendendo i dati più bilanciati per l'addestramento del modello, *l'undersampling* è, invece, la tecnica che permette di ridurre il numero di esempi della classe maggioritaria per bilanciare il *dataset*. Si veda R. Mohammed *et al.*, *Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results,* in *2020 11th International Conference on Information and Communication Systems,* 7 Settembre 2020.

[94] R. I. Júnior, *et al.*, *Data Balancing for Mitigating Sampling Bias in Machine Learning,* in *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, 1204-1212.

[95] G. D'Acquisto, *Dati sintetici: cosa sono, le applicazioni e i rischi da gestire,* in *Agenda Digitale,* 6 Maggio 2024.

consente di riequilibrare sbilanciamenti interni ai *dataset,* generando esempi aggiuntivi per categorie sottorappresentate e contribuendo, di conseguenza, a ridurre le forme di *bias* nei sistemi di IA[96].

In tale prospettiva, l'obiettivo della *fairness* algoritmica non sembra coincidere con una mera neutralità formale, ma si avvicina alla logica dell'eguaglianza sostanziale. L'esigenza di costruire *dataset* rappresentativi, di prestare particolare attenzione ai gruppi sottorappresentati e di introdurre meccanismi correttivi capaci di compensare squilibri preesistenti richiama infatti la funzione promozionale attribuita dall'art. 3, secondo comma, Cost. ai pubblici poteri. Come si è accennato, anche il monitoraggio umano costante dei risultati (nell'ottica della c.d. *humancentric AI*), audit regolari e processi di validazione sono cruciali per identificare e correggere i bias nei sistemi di IA. Si evidenzia che l'istituzione di una sorta di dipartimento dedicato al controllo della qualità degli algoritmi all'interno degli ospedali garantirebbe un monitoraggio continuo, l'identificazione dei *bias* e gli aggiornamenti necessari a un'efficace assistenza ai pazienti[97]. In tal senso, l'approccio *human-in-the-loop*[98] prevede che i sistemi di IA siano progettati in modo da consentire un intervento umano costante nelle attività di controllo e supervisione, garantendo così un utilizzo appropriato della tecnologia. Anche l'*Ai Act* richiede che gli operatori siano messi nelle condizioni di comprendere le capacità e i limiti dell'algoritmo, monitorarne le prestazioni e intervenire quando necessario[99]. Tale configurazione aiuta a evitare che l'utente riponga un'eccessiva fiducia nei risultati automatizzati, limitando così il rischio di *bias* e favorendo decisioni più corrette ed equilibrate[100].

È, infine, fondamentale educare al tema medici e pazienti, così da promuovere sistemi sanitari globalmente più equi. La consapevolezza dei *bias* consente agli operatori di valutare criticamente gli *output* dell'IA e prendere decisioni più informate. Anche i pazienti, se consapevoli dei *bias* dell'IA, possono dialogare in modo più attivo con gli operatori sanitari e avere un ruolo più partecipe nel percorso clinico e nell'alleanza terapeutica[101]. Questo *empowerment* favorisce un'assistenza centrata sulla persona e garantisce che anche le preferenze individuali e le circostanze specifiche siano considerate[102]. Non a caso, l'art. 4 dell'*Ai Act*[103] introduce un obbligo specifico che mira a garantire che il personale coinvolto nell'uso dei sistemi di IA (fornitori e *deployer*) possieda un sufficiente livello di *AI literacy* e consapevolezza su rischi e opportunità connessi ai sistemi di IA[104]. Le suddette strategie di mitigazione si configurano, in un certo senso, anche come strumenti di attuazione dell'eguaglianza sostanziale quale principio che informa lo Stato democratico. Proprio ciò giustificherebbe, nel contesto della sanità pubblica, anche un intervento statale in tal senso, e non solo di

---

[96] J. Blackman, E. Burgio, *AI bias: come i pregiudizi influenzano l'intelligenza artificiale,* in *Agenda Digitale,* 12 marzo 2025.

[97] D. Ueda *et al.*, *Fairness of artificial intelligence in healthcare: review and recommendations*, in *Japanese Journal of Radiology*, 42, 2024, 3 - 15.

[98] D. Martire, *Human in the loop. L'essere umano come fattore condizionante della – o condizionato dalla – intelligenza artificiale,* in *Rivista Italiana di Informatica e Diritto*, 2, 2024, 467- 481.

[99] Si riporta in tal senso l'esempio di *OpenAI*, che, nell'intento di contenere gli abusi, ha illustrato come integra la supervisione umana attraverso il proprio processo di allineamento, con l'obiettivo di garantire che i sistemi di IA operino secondo gli scopi previsti e nel rispetto dei principi etici di riferimento.

[100] J. Blackman, E. Burgio, *AI bias: come i pregiudizi influenzano l'intelligenza artificiale,* cit.

[101] Sull'evoluzione della relazione medico-paziente si vedano S. Nanni, *eHealth: evoluzione del rapporto medico-paziente e questioni giuridiche,* in *Agenda Digitale*, 31 Luglio 2020; S. Canestrari, *"Una buona legge buona": la l. n. 219 del 2017 e la relazione medico-paziente*, in *Diritto e Salute,* 4, 2018, 51-60; G. Vicarelli, *Le trasformazioni del rapporto medico-paziente in Italia: una analisi processuale,* in *Scuola di Etica Medica*, 2019.

[102] D. Ueda, *et al., Fairness of artificial intelligence in healthcare: review and recommendations*, cit.

[103] Cfr. Regolamento (UE) 2024/1689, art. 4.

[104] T. Ricci, *AI Act, partiti i divieti: ecco che devono fare le aziende UE,* in *Agenda Digitale*, 3 Febbraio 2025.

regolazione privata, nella progettazione dei sistemi.
Come si è illustrato già *supra,* recenti studi hanno evidenziato una maggiore fiducia da parte degli utenti nei confronti di assistenti robotici, questo è spiegabile in forza di una maggiore potenziale imparzialità del sistema di IA nell'adottare decisioni. Ciò grazie al fatto che l'"adozione di tecnologie di supporto decisionale clinico basate sull'IA può portare a una rimozione parziale dei *bias* impliciti e inconsci umani[105] e a una maggiore standardizzazione dei processi decisionali. I sistemi di IA sono infatti, a differenza dell'uomo, in grado di trattare grandi quantità di dati clinici in modo coerente, evitando variabilità dovute a fattori soggettivi degli operatori sanitari e di assumere decisioni basate su evidenze cliniche (c.d. *data-driven*) che permettono una maggiore uniformità nel trattamento dei pazienti[106].
Per completezza, va osservato, inoltre, che agenti intelligenti possono anche essere impiegati per mitigare i *bias* cognitivi e sociali nelle interazioni umane. La *Fair Proxy Communication* (FPC) rappresenta, ad esempio, un'applicazione pionieristica della robotica sociale basata su tecnologie di telepresenza, con l'obiettivo primario di esplorare i meccanismi della cognizione sociale implicita in ambito neuropsicologico. Un'applicazione esemplare di tale tecnologia consiste nell'inserimento di un robot neutrale durante un colloquio di lavoro, al fine di tutelare l'anonimato e ridurre i pregiudizi legati all'identità sociale. Non si può tuttavia escludere che tale modello potrà in futuro essere applicato e portare benefici anche al settore sanitario[107].

7. Sistemi di IA quale veicolo di inclusione sociale. Nel tentativo di esplorare, senza alcuna pretesa di esaustività, gli indiscutibili benefici derivanti da applicazioni di sistemi di IA nel settore sanitario, è opportuno specificare che in tale settore si registrano iniziative pionieristiche promettenti pensate per abbattere barriere sociali e culturali che limitano l'accesso equo ai servizi di cura. Ciò, soprattutto grazie alla progressiva diffusione di: sistemi diagnostici a distanza (*Clinical Decision Support System*, CDSS); strumenti intelligenti per il monitoraggio dello stato di salute, che garantiscono una più effettiva attuazione del diritto alla salute anche a beneficio di pazienti che si trovano in aree geograficamente remote o da parte di persone affette da disabilità rilevanti; assistenti vocali e sistemi di riconoscimento vocale a supporto delle persone non vedenti; piattaforme di educazione sanitaria personalizzate.
L'OMS, nel rapporto *Ethics and Governance of Artificial Intelligence for Health*, sintetizzando i principi etici fondamentali cui dovrebbe sempre ispirarsi l'impiego dell'IA in ambito sanitario pone il *focus* su inclusività ed equità e sulla necessità di sistemi progettati in modo da favorire un accesso ampio e paritario a tali tecnologie, evitando la codificazione di pregiudizi ai danni di gruppi già emarginati o vulnerabili[108].
Con riferimento specifico al contesto normativo nazionale, va menzionata la recente l. n. 132/2025. Il testo, composto da 26 articoli, dedica particolare attenzione all'inclusione delle persone con disabilità. L'art. 7, c. 4, in particolare, specifica che la legge «promuove lo sviluppo, lo studio e la diffusione di sistemi di IA che migliorano le condizioni di vita delle

[105] C. Hobson *et al.*, *Removing Data Bias from AI and Machine Learning Tools in Healthcare White Paper,* in *Himss Iowa Chapter,* 25 Febbraio 2021.
[106] F. Hasanzadeh *et al.*, *Bias recognition and mitigation strategies in artificial intelligence healthcare applications*, in *Npj Digital Medicine,* 154, 2025.
[107] J. Skewes *et al., Social robotics and the modulation of social perception and bias,* in *Philosophical Transactions B*, 374, 2019, 2; J. Seibt, C. Vestergaard, *Fair Proxy Communication: Using Social Robots to Modify the Mechanisms of Implicit Social Cognition*, in *Research Ideas and Outcomes*, 3, 2018.
[108] In tal senso sono le dichiarazioni di T.A. Ghebreyesus, Direttore generale dell'OMS, in merito al rapporto, OMS, *Ethics and Governance of Artificial Intelligence for Health*, 2021.

persone con disabilità, agevolano l'accessibilità, l'autonomia, la sicurezza e i processi di inclusione sociale delle medesime persone, anche ai fini dell'elaborazione del progetto di vita…». Tale previsione si configura come un importante riconoscimento del ruolo trasformativo dell'IA nel garantire diritti di cittadinanza attiva e nel rafforzare le politiche di coesione sociale rivolte alle fasce più fragili della popolazione ciò anche in linea con la Convenzione ONU sui diritti delle persone con disabilità[109], che specifica il percorso che gli Stati aderenti devono seguire per garantire l'eguaglianza e l'inclusione sociale di tutti i soggetti con disabilità anche nel contesto salute e che all'art. 26 «Abilitazione e riabilitazione», in particolare al c. 3, mette l'accento sulla promozione dell'uso di tecnologie per l'inclusione[110]. Ponendo nuovamente il *focus* sui sistemi di IA emozionale, si sottolinea che l'*affective computing* fornisce vantaggi sostanziali in ambito medico, tra cui il miglioramento dell'assistenza al paziente e della comunicazione medico-paziente. Le applicazioni principali più vantaggiose si riscontrano nella diagnosi della depressione, dove l'*affective computing* aumenta l'oggettività e l'accuratezza diagnostica attraverso modelli di classificazione che analizzano e combinano diversi tipi di dati (audio, EEG e facciali); nell'identificazione di casi di Disturbi dello Spettro Autistico, grazie a modelli di classificazione delle emozioni, e nell'intervento tramite tecnologie come la realtà virtuale e sensori multimodali; e nel riconoscimento e classificazione dei livelli di dolore[111]. Altre applicazioni comprendono il supporto nella diagnosi e nel monitoraggio del disturbo bipolare, la compagnia e il monitoraggio degli anziani, lo sviluppo di piattaforme mediche intelligenti per il benessere emotivo tramite l'analisi di segnali fisiologici. I limiti di queste tecnologie non sono intrinseci, ma, come si è visto, derivano dalla complessità della psiche umana e del dato emotivo. Ciononostante, l'identificazione delle emozioni, sebbene inevitabilmente imperfetta, presenta vantaggi concreti rilevanti. Anche soggetti che hanno difficoltà a riconoscere autonomamente le emozioni, come le persone non vedenti, possono trarne grande beneficio[112]. Si noti che i ricercatori dell'Università di Stanford stanno già studiando metodi per dotare individui nello spettro autistico di occhiali intelligenti in grado di riconoscere le emozioni di chi li circonda. Altri studi recenti mostrano un crescente interesse per l'integrazione di robot sociali nella pratica clinica rivolta a persone con Disturbo dello Spettro Autistico (ASD). Questi soggetti potrebbero trarre beneficio dall'interazione con robot sanitari, in particolare per stimolare la loro comunicazione e interazione sociale. Secondo queste ricerche, i pazienti in questione riescono a interagire più facilmente con un partner robotico rispetto a uno umano. Gli studi suggeriscono che ciò dipenda principalmente da tali caratteristiche dei robot: possono agire come strumenti motivazionali e sono agenti semplificati il cui comportamento è più prevedibile rispetto a quello umano[113]. In effetti, già diversi anni fa, i risultati di uno studio pilota sul comportamento di cinque bambini con autismo hanno mostrato benefici sui pazienti, quando esposti a semplici moduli di interazione uomo-robot (HRI) col robot umanoide NAO. Gli scienziati hanno osservato che l'aspetto umanoide di NAO, la sua

[109] La convenzione delle Nazioni Unite sui diritti delle persone con disabilità (CDPD) è stata adottata nel 2006 e ratificata dall'Italia con l. n. 18/2009.

[110] Per un approfondimento sulle tecnologie *AI based* per l'inclusione delle persone con disabilità si veda F. Moioli, *Intelligenza artificiale per l'inclusione sociale: tecnologie e step necessari,* in *Agenda Digitale*, 29 Giugno 2021.

[111] J. Apablaza, S. Cano, *Affective Computing from Digital Health: A literature review,* cit.

[112] Si veda sul punto S. Ashmad Ahemed *et a., Voice Assisted Facial Emotion Recognition System For Blind Peoples With Tensorflow Model,* in *IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS),* Bhopal, India, 2024, pp. 1- 4.

[113] M. Dubois-Sage, *et al.*, *People with Autism Spectrum Disorder Could Interact More Easily with a Robot than with a Human: Reasons and Limits,* in *Behavioral sciences*, 14, 2, 2024; R. Vagnetti *et al., Social Robots: A Promising Tool to Support People with Autism. A Systematic Review of Recent Research and Critical Analysis from the Clinical Perspective,* in *Review Journal of Autism and Developmental Disorders,* 2024, 1- 25.

capacità di sbattere le palpebre, parlare e riprodurre musica, unita alla semplicità dei moduli di interazione, erano in grado di suscitare l'interesse dei bambini a comunicare. Questo ha portato a una riduzione significativa[114] delle caratteristiche autistiche nei comportamenti comunicativi rispetto al contesto scolastico abituale[115]. Le applicazioni finora prospettate mostrano come l'IA ha il potenziale per rappresentare uno strumento di inclusione e di rimozione degli ostacoli che limitano l'effettiva partecipazione delle persone con disabilità alla vita sociale in attuazione degli artt. 9 e 26 della Convenzione ONU sui diritti delle persone con disabilità e dell'art. 26 della Carta di Nizza[116]. Da ciò discende anche la necessaria dimensione pubblicistica nella progettazione e nell'implementazione di queste tecnologie[117], atteso anche che la loro diffusione può concorrere all'attuazione degli obblighi positivi gravanti in capo allo stato *ex* c. 2 dell'art. 3 Cost.

8. Riflessioni conclusive. La discriminazione algoritmica si colloca fra le principali sfide che le nuove tecnologie *AI based* pongono alla tutela dei diritti fondamentali, imponendo una rinnovata interpretazione del principio di uguaglianza. Si è già avuto modo *supra* di esplorare le problematiche relative all'utilizzo dei sistemi di IA nel campo medico, in particolare in riferimento ai possibili esiti discriminatori che violano il principio di uguaglianza formale sancito dal primo comma dell'articolo 3 Cost. Risulta altrettanto fondamentale evidenziare gli aspetti rilevanti legati al principio di uguaglianza sostanziale, di cui al secondo comma dello stesso articolo. Va detto, infatti, che un sistema di IA che sistematicamente sottostima i bisogni di cura delle minoranze non viola solo il divieto di discriminazione di cui all'art. 3, c. 1 Cost., ma compromette l'adempimento del progetto costituzionale sotteso dall'art. 3, c. 2, secondo cui lo Stato ha un obbligo positivo di rimuovere gli ostacoli che impediscono l'accesso equo alle cure. La questione dei *bias* algoritmici non investe, dunque, esclusivamente il principio di eguaglianza formale, ma coinvolge la dimensione promozionale e redistributiva dell'eguaglianza sostanziale nel suo saldo connubio col principio solidaristico. Come si è cercato di evidenziare, l'implementazione di sistemi di IA può, infatti, costituire tanto uno strumento di riproduzione delle disuguaglianze esistenti quanto un mezzo idoneo alla loro riduzione, a seconda delle modalità con cui la tecnologia viene progettata, addestrata e governata.

Da un lato, è evidente, poi, che i sistemi di IA, e in particolare i sistemi di IA emozionale, possono contribuire positivamente alla cura e all'assistenza medica di gruppi vulnerabili, come persone con disabilità, persone anziane, soprattutto se residenti in zone geografiche remote, o persone con disturbi dello spettro autistico. Analogamente, anche soggetti con problemi psicologici possono trarre vantaggio dalle cosiddette piattaforme empatiche di IA, dai *chatbot* e dagli assistenti virtuali in grado di interagire emotivamente a un livello base, come *Woebot* per le malattie psichiche e *Tess* per le persone non udenti, che offrono un primo

[114] In particolare, quattro dei cinque bambini hanno mostrato una diminuzione dei comportamenti autistici (nell'ambito della comunicazione), influenzando positivamente il loro comportamento comunicativo.

[115] S. Shamsuddin *et al.*, *Humanoid Robot NAO Interacting with Autistic Children of Moderately Impaired Intelligence to Augment Communication Skills,* in *Procedia Engineering*, 2012,1533-1538.

[116] G. Quinn, *The impact of new technologies on the rights of persons with disabilities*, A/HRC/49/52, Report of the Special Rapporteur on the rights of persons with disabilities, United Nations, in *The Human Rights Council at its 49th session*, 2021.

[117] Sull'utilizzo di robot assistivi e tutela della salute in chiave pubblicistica, nonché sulla tecnologia quale veicolo per colmare i gap anche territoriali negli standard di cura cfr. E. Ferioli, *Il diritto alla salute alla prova della Digital Health*, in *Politiche Sociali*, 2, 2023, 207-226, nonché E. Ferioli, *Digitalizzazione, Intelligenza Artificiale e Robot nella tutela della salute,* cit.

supporto rispondendo agli *input* degli utenti con empatia pre-programmata[118]. D'altro canto, sebbene i sistemi di IA possano offrire importanti strumenti di supporto per le persone con disabilità – migliorando l'accessibilità, la comunicazione e l'autonomia personale – anche questa categoria applicativa può presentare criticità strutturali. Si pensi che nel dibattito sull'*AI Act*, alcuni hanno sostenuto un divieto totale del riconoscimento delle emozioni nella regolamentazione, opponendosi anche alla possibilità di un'eccezione per l'uso di queste tecnologie a supporto delle persone con disabilità o neurodivergenti. In particolare, al di là del rischio che l'IA possa commettere errori nell'interpretare le espressioni di persone con autismo o con disabilità o con patologie che influenzano i movimenti facciali, si pone la questione del rafforzamento dello stigma nei loro confronti, in quanto tali tecnologie possono promuovere stereotipi abilisti[119]. Esiste il rischio concreto che tali modelli, se progettati senza un'adeguata comprensione delle dinamiche sociali e culturali, finiscano per rafforzare e perpetuare l'ancestrale stigma nei confronti della disabilità. Ad esempio, l'adozione di sistemi di riconoscimento delle emozioni in persone con disturbi neurodivergenti o con limitazioni motorie del volto può risultare non solo inefficace, ma anche discriminatoria, in quanto si basa su modelli espressivi *standard* che non riflettono la pluralità delle manifestazioni umane. La stessa idea implicita che le persone con disabilità abbiano necessariamente bisogno dell'assistenza costante di tecnologie intelligenti può contribuire a rappresentazioni paternalistiche o medicalizzanti della disabilità stessa.

Il *Center for Democracy & Technology* (CDT) ha pubblicato un rapporto che evidenzia come le persone con disabilità siano particolarmente a rischio di discriminazione quando interagiscono con sistemi IA e algoritmici. Molti di questi strumenti sono addestrati su *pattern* tipici nei *dataset*, che non rappresentano adeguatamente le esperienze delle persone con disabilità, ma forniscono meri ritratti parziali o imprecisi[120]. Una recente ricerca ha, infatti, evidenziato che modelli di IA generativa di immagini, come *Stable Diffusion XL* e *DALL-E* 3, mostrano *bias* nella rappresentazione delle persone con disabilità[121], spesso raffigurandole come anziane, tristi o esclusivamente su sedie a rotelle manuali, perpetuando stereotipi e immagini negative della disabilità[122]. Da qui la necessità che i sistemi di IA siano progettati *by design* e verificati per garantire che non amplifichino stigmi sulle persone con disabilità, in ossequio a dannosi paradigmi abilisti[123]. Pericoli di stigmatizzazione e ulteriore marginalizzazione sono associati, infatti, allo sviluppo e all'uso di tecnologie assistive soprattutto se i dispositivi sono percepiti come segnali ben visibili e vistosi di *deficit* o diversità[124].

---

118 L. Spytska, *The use of artificial intelligence in psychotherapy: development of intelligent therapeutic systems,* cit.; R. Fulmer *et al., Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial*, in *JMIR Mental Health,* 5, 4, 2018.

119 M. Martorana, R. Savella, *IA e riconoscimento delle emozioni: rischi e possibili vantaggi,* cit.

120 Si rimanda a *report* quali A. Aboulafia, H. Claypoo, *Building A Disability-Inclusive AI Ecosystem;* in *CDT e AAPD,* 2025; T. Krupiy, *Disability Discrimination in the Digital Realm: How the ICRPD Applies to Artificial Intelligence Decision-Making Processes and Helps in Determining the State of International Human Rights Law,* in *Human Rights Law Review,* 23, 3, 2023.

121 Cfr. Y. Tevissen, *Disability Representations: Finding Biases in Automatic Image Generation*, in *Computer Vision and Pattern Recognition*, 2024; R. Manzoor *et al.*, *Out of dataset, out of algorithm, out of mind: a critical evaluation of AI bias against disabled people,* in *AI & Society,* 40, 2025, 3941 - 3951.

122 Si è notato che anche strumenti di analisi del *sentiment* tendono a classificare termini legati alla disabilità (*e.g. blind*) come negativi o tossici, indipendentemente dal contesto. Si rimanda sul punto a P. Narayanan Venkit, *Automated Ableism: An Exploration of Explicit Disability Biases in Sentiment and Toxicity Analysis Models,* in *Computation and Language,* 18 luglio 2023.

123 American Foundation for the Blind, *Guiding Principles for More Disability-Inclusive AI,* in *AFB.org*.

124 Era questo il caso di tecnologie progettate per persone con disturbo dello spettro autistico (ASD). Si è sottolineato che l'uso di tali tecnologie può portare alla stigmatizzazione degli utenti, soprattutto se i dispositivi

Un altro spunto di riflessione riguarda il fatto che sussistono oggi ostacoli nell'accesso ai servizi base di sanità digitale e, a maggior ragione, a quelli tecnologicamente più avanzati. I robot sociali, ad esempio, sono impiegati solo in strutture pubbliche in modo sporadico o all'interno di progetti pilota. Anche qualora questi dispositivi venissero commercializzati, i loro costi elevati ne precluderebbero comunque l'accesso a fasce ampie della popolazione bisognosa (c.d. *robotics divide*[125]). Anche l'accesso ai sistemi di IA integrati in *chatbot* e assistenti per la salute mentale non può dirsi, allo stato attuale, equo. Parti della popolazione, specie gli anziani, non possiedono un'alfabetizzazione digitale sufficiente per utilizzare efficacemente e consapevolmente tali strumenti (c.d. *AI divide*[126]). In Italia, si registra ancora un ritardo sia nelle competenze digitali avanzate sia in quelle di base, tra i cittadini e tra i dipendenti pubblici in sanità, con un *gap* significativo rispetto agli altri paesi europei e anche un *gap* interno a sfavore delle regioni meridionali[127]. Se è vero che l'analisi algoritmica dei dati può produrre risultati capaci di mitigare e correggere atteggiamenti pregiudizievoli, favorendo i gruppi sociali più deboli, è però altrettanto evidente che questa sfida non può essere affrontata senza un intervento umano. Se le condizioni di partenza sono inique, affidare alle macchine il compito di compiere scelte giuste è inutile: esse elaboreranno semplicemente i dati parziali forniti loro dagli esseri umani, riflettendo la realtà percepita. È chiaro, dunque, che se a una parte significativa dell'umanità sono ancora negati i servizi sanitari minimi, tale disuguaglianza non potrà che ampliarsi se solo pochi potranno scegliere non solo di curarsi, ma anche di potenziare la propria salute[128].

Abstract. Il contributo analizza le sfide e le opportunità poste dal crescente impiego di sistemi di IA – con particolare attenzione ai sistemi di *emotional AI* – nel settore sanitario e il loro impatto sul principio di eguaglianza. Passando in rassegna le principali fonti e tipologie di *bias*, un focus particolare viene posto sui rischi di discriminazione algoritmica associati alle suddette tecnologie e su possibili strategie di mitigazione dei *bias*. L'analisi considera, inoltre, i potenziali benefici che i sistemi di IA offrono in termini di democratizzazione dei processi decisionali e di promozione dell'inclusione sociale.

Abstract. The paper examines the challenges and opportunities arising from the increasing use of AI systems – particularly emotional AI – in the healthcare, and their impact on the principle of equality. By reviewing the main sources and types of bias, special attention is posed to the risks of algorithmic discrimination associated with these technologies and to possible strategies for bias mitigation. The analysis also considers the potential benefits of AI systems in terms of democratizing decision-making processes and fostering social inclusion.

Parole chiave. IA emozionale – Diritto alla salute – Diritto antidiscriminatorio – *Bias* – Principio di eguaglianza.

sono vistosi. Adottando un approccio centrato sull'utente e sensibile alle implicazioni sociali, è opportuno che tali dispositivi siano progettati non solo per essere funzionali, ma anche per essere esteticamente discreti e socialmente accettabili. Cfr. F. O'Brolcháin, B. Gordijn, *Risks of Stigmatisation Resulting from Assistive Technologies for Persons with Autism Spectrum Disorder,* in *Technologies,* 6, 27, 2018.

[125] A. López Peláez (a cura di) *The Robotics Divide. A New Frontier,* in *the 21st Century?*, London, 2014.

[126] C. Casonato, *Per una intelligenza artificiale costituzionalmente orientata*, cit., 143.

[127] E. Sorrentino *et al., Le sfide degli enti locali: tra PNRR e gap digitali,* in *Federalismi.it,* 21, 2023, 160.

[128] U. Ruffolo, A. Amidei, *Intelligenza artificiale, biotecnologie e potenziamento: verso nuovo diritti della persona?,* in U. Ruffolo (a cura di), *XXVI lezioni di diritto dell'intelligenza artificiale,* Torino, 2021, 108.

Key words. Emotional AI – Right to health – Antidiscrimination Law – Bias – Principle of equality.