D2.1 Relazione sulla parzialità

Cofinanziato dall'Unione europea

Indice

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

Introduzione

Parte del progetto AEQUITAS consiste nella creazione di un database sui pregiudizi di genere e razziali nelle applicazioni mediche dell'intelligenza artificiale (IA), con particolare attenzione a tre malattie: malattie cardiovascolari, diabete e depressione.

Per portare a termine questo compito, i partner del consorzio devono innanzitutto raccogliere una serie di fonti sui pregiudizi sopra elencati. L'Ospedale universitario di Colonia (Universitätsklinikum Köln, UKK), in qualità di responsabile del compito e esperto del settore, ha organizzato l'attività di raccolta delle informazioni e ha fornito il modello di mappatura che i partner hanno utilizzato per mappare le fonti, garantendo che le informazioni rilevanti potessero essere facilmente trasferite al database.

Il presente rapporto illustra i fondamenti teorici e scientifici che hanno guidato la selezione dell'attività di raccolta e del modello di mappatura, accompagnati da casi di studio che mostrano i diversi tipi di pregiudizi; le implicazioni politiche dei pregiudizi indotti dall'IA biomedica sui diritti tutelati dalla Carta dei diritti fondamentali dell'Unione europea; una descrizione dell'attività di raccolta dati; il modello di mappatura; un elenco delle fonti raccolte dai partner di AEQUITAS; e altro materiale di supporto. Il resto del rapporto è strutturato come segue:

In primo luogo, presentiamo il contesto teorico relativo al nostro lavoro in un'introduzione sulle applicazioni mediche dell'IA e sul concetto di pregiudizio nei sistemi informatici e in medicina. Iniziamo concentrandoci sulla medicina, presentando in primo luogo come i pregiudizi razziali e di genere si manifestano nell'assistenza medica e, in secondo luogo, come le questioni morali che sorgono nella pratica della medicina e nella ricerca biomedica sono affrontate dalla bioetica, offrendo una breve introduzione ai quattro principi della bioetica (autonomia, non maleficenza, beneficenza e giustizia).

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

Successivamente, passiamo al dominio dell'IA, presentando i tipi di pregiudizio che possono essere osservati nei sistemi di IA così come si manifestano nella pipeline di Machine Learning e Intelligenza Artificiale (ML/AI). Ogni tipo di pregiudizio è accompagnato da esempi e da un caso di studio sui pregiudizi di genere e razziali e sul loro impatto a livello sociale in relazione alle tre malattie oggetto del progetto AEQUITAS (cardiovascolari, diabete e depressione), tratti da fonti raccolte dai partner AEQUITAS dopo il completamento della fase T2.2. Quando ciò non è stato possibile perché il materiale raccolto non dimostrava chiaramente il tipo specifico di pregiudizio dell'IA in esame, è stato presentato un caso alternativo proveniente da un altro settore medico facilmente generalizzabile alle malattie target di AEQUITAS. Le descrizioni dei casi di studio, insieme alle fonti raccolte, attingono ad ulteriori risorse scientifiche, se necessario, a loro sostegno.

Infine, nella sezione "Implicazioni politiche", viene dimostrato come i vari tipi di pregiudizi dell'IA influiscano sui diritti fondamentali tutelati dalla Carta dell'UE, in particolare i precetti della dignità umana, dell'uguaglianza davanti alla legge, della non discriminazione, nonché il diritto all'integrità della persona, il diritto all'assistenza sanitaria, la protezione dei dati e il diritto a un ricorso effettivo, concludendo con le garanzie che possono essere messe in atto nelle valutazioni di conformità, nel monitoraggio post-commercializzazione e negli appalti pubblici.

La relazione si conclude con i riferimenti bibliografici e le seguenti appendici:

Appendice 1: Raccolta delle fonti e metodo di mappatura, che contiene il modello di mappatura e descrive il processo di raccolta, mappatura e valutazione delle informazioni condotto durante le attività T2.1 e T2.2.

Appendice 2: contiene materiale di supporto per le attività T2.1 e T2.2, ovvero le diapositive delle riunioni dei partner che descrivono il processo, presentate da UKK.

Appendice 3: Elenco delle fonti raccolte dai partner AEQUITAS.

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

Applicazioni mediche dell'IA

L'ascesa delle applicazioni di IA negli ultimi anni ha avuto un forte impatto sulla medicina, compresa l'acquisizione di dati digitalizzati, l'apprendimento automatico e l'infrastruttura informatica (Yu et al., 2018). In particolare, l'introduzione di algoritmi di deep learning in settori quali la visione artificiale e l'elaborazione del linguaggio naturale ha rivoluzionato le applicazioni informatiche in radiologia, patologia, cardiologia, diabetologia, psichiatria, oncologia, ecc. (Esteva et al., 2019; Koteluk et al., 2021; Rajpurkar et al., 2022; Gou et al., 2024). L'Organizzazione Mondiale della Sanità (OMS) elenca i seguenti ambiti di applicazione dei sistemi di IA nell'assistenza sanitaria: diagnosi e diagnosi basata sulla previsione, assistenza clinica, ricerca e sviluppo di farmaci, gestione e pianificazione dei sistemi sanitari, sanità pubblica e sorveglianza della sanità pubblica, promozione della salute, prevenzione delle malattie, sorveglianza basata sulla previsione, preparazione alle emergenze e risposta alle epidemie (Organizzazione Mondiale della Sanità, 2021).

Tuttavia, l'avvento delle applicazioni di IA in medicina comporta una serie di sfide, quali difficoltà di implementazione, tra cui la fiducia nei modelli e le limitazioni dei dati, questioni di responsabilità, che includono sfide normative e corretta attribuzione delle responsabilità, e la garanzia di equità attraverso l'uso etico dei dati, la distribuzione equa dei benefici e l'individuazione e la mitigazione dei pregiudizi (Rajpurkar et al., 2022).

Il progetto AEQUITAS si concentra sui casi di pregiudizio di genere e razziale nelle malattie cardiovascolari, nel diabete e nella depressione. Le applicazioni mediche dell'IA supportano la cura cardiovascolare attraverso il supporto alle decisioni cliniche, la telemedicina, la valutazione dei rischi, la terapia personalizzata, l'analisi predittiva e il monitoraggio remoto (Bernstein et al., 2025; Naskar et al., 2025), migliorano la

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

gestione del diabete (compreso il monitoraggio dei pazienti e l'autogestione), la diagnosi, il trattamento e la prevenzione (Contreras & Vehi, 2018; Khalifa & Albadawy, 2024; Naskar et al., 2025; Sheng et al., 2024). Per quanto riguarda la depressione, sono coinvolti nello screening, nella diagnosi e nel trattamento (Alhuwaydi, 2024) con un'attenzione particolare alla rilevazione e allo screening con l'uso di modelli linguistici di grandi dimensioni (LLM) (Cao et al., 2025; Kumari et al., 2025; Mao et al., 2023; Wang et al., 2025). In tutte le aree sopra citate esistono sfide legate al bias, ad esempio per quanto riguarda le malattie cardiovascolari, il diabete e la depressione, cfr. (van Assen et al. 2024), (Cronjé et al. 2023), (Dang et al. 2024), rispettivamente.

Le sfide come la distorsione, sia in medicina che nell'IA, vengono affrontate attraverso una combinazione di bioetica ed etica dell'IA. Nella sezione seguente, forniamo una breve panoramica di questi due tipi di domini di etica applicata che sono serviti come base teorica e scientifica per lo sviluppo del modello di mappatura della distorsione.

Etica e pregiudizi nella medicina e nell'IA

Bioetica e pregiudizi in medicina

I pregiudizi in medicina sono ben documentati: si veda, ad esempio, (Hammond et al. 2021) per quanto riguarda i pregiudizi cognitivi, che consistono in errori sistematici di pensiero dovuti a limitazioni dell'elaborazione umana o a modelli mentali inappropriati, e (FitzGerald e Hurst 2017) per i pregiudizi impliciti che coinvolgono associazioni al di fuori della consapevolezza cosciente che portano a una valutazione negativa di una persona sulla base di caratteristiche irrilevanti come la razza o il sesso.

I pregiudizi razziali in medicina sono stati studiati approfonditamente negli Stati Uniti, dove è stato documentato che gli afroamericani, così come altri gruppi minoritari, ricevono meno cure mediche e di qualità inferiore, ottengono trattamenti meno

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

aggressivi, subiscono tassi più bassi di trattamenti chirurgici e ricevono meno referenze da specialisti rispetto alle persone bianche (Bowser, 2001; Williams & Wyatt, 2015).

Il pregiudizio di genere può essere attribuito alla cecità di genere e ai preconcetti stereotipati su uomini e donne (Hamberg, 2008), oltre che a una generale mancanza di conoscenza del funzionamento del corpo femminile e delle sue differenze biologiche rispetto al corpo maschile. Ad esempio, le donne di età superiore ai 50 anni in condizioni critiche avevano meno probabilità rispetto agli uomini in condizioni critiche di essere ricoverate in un'unità di terapia intensiva (ICU) (Bierman, 2007), e persino i modelli murini maschi sono complessivamente più rappresentati rispetto ai modelli femminili nella ricerca biomedica di base, preclinica e chirurgica (Yoon et al., 2014).

È inoltre importante notare che le persone LGBT+ subiscono discriminazioni in termini di accesso all'assistenza sanitaria e sono soggette a stereotipi che non riguardano la popolazione eterosessuale. Questi fattori sociali e culturali perpetuano la discriminazione e hanno un impatto sulla salute. Ad esempio, uno studio condotto negli Stati Uniti, basato sui dati del National Health Interview Survey (NHIS) 2013-14, ha rilevato che gli adulti LGB hanno riportato livelli più elevati di cattiva salute, limitazioni funzionali, grave disagio psicologico e difficoltà a permettersi l'assistenza sanitaria rispetto ai loro omologhi eterosessuali. Queste disuguaglianze sono determinate dallo stress dei gruppi minoritari e dalla marginalizzazione sociale multifattoriale (Liu et al., 2023).

D'altra parte, la medicina come disciplina è soggetta a elevati standard etici dall'antichità ai giorni nostri (Baker & McCullough, 2008). Da secoli, la società si aspetta che i medici seguano le regole etiche di responsabilità professionale stabilite dagli standard della loro professione, come dimostrano le norme professionali che vanno dal Giuramento di Ippocrate del 400 a.C. (Miles, 2005) alle Dichiarazioni di Ginevra e di Helsinki (Tröhler, 2008). Come sottolineato da (Vevaina et al., 1993), i medici sono tenuti a conformarsi al codice etico della loro professione in virtù

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

dell'investimento che la società fa nella loro formazione (in termini monetari e nell'uso dei suoi membri come materiale didattico durante la formazione e la carriera del medico) e del monopolio virtuale che la loro professione si vede garantire attraverso il rilascio delle licenze.

L'etica biomedica (o bioetica) è un dominio dell'etica pratica (o applicata) che affronta le questioni morali che sorgono nella pratica della medicina e nella ricerca biomedica (Vevaina et al., 1993). Al centro dell'etica biomedica ci sono i quattro principi definiti da Beauchamp e Childress (Beauchamp & Childress, 2019):

1. Autonomia: rispetto delle capacità decisionali delle persone autonome. Due condizioni generali sono essenziali per l'autonomia: la libertà, che si manifesta come indipendenza da influenze controllanti, e l'azione, ovvero la capacità di agire intenzionalmente.

2. Non maleficenza: evitare di causare danni.

3. Beneficenza: intraprendere azioni positive per aiutare gli altri, in particolare prevenendo il male o il danno, rimuovendo il male o il danno e promuovendo il bene.

4. Giustizia: distribuire equamente benefici, rischi e costi. La giustizia è interpretata come un trattamento equo, imparziale e appropriato per gli individui e i gruppi, date le numerose disparità nell'assistenza sanitaria e nella ricerca basate su razza, etnia, genere e status sociale.

Etica e pregiudizi dell'IA

L'introduzione dell'IA e il rapido sviluppo delle sue applicazioni hanno sollevato una serie di questioni etiche (Christoforaki & Beyan, 2022), tra cui spiccano il pregiudizio e la discriminazione.

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

Pertanto, l'etica dell'IA è stata sviluppata come un dominio dell'etica pratica (o applicata) che comprende "un insieme di valori, principi e tecniche che utilizzano standard ampiamente accettati di giusto e sbagliato per guidare la condotta morale nello sviluppo e nell'uso delle tecnologie di IA" (Leslie, 2019, p. 3).

L'etica dell'IA attinge sia dalla bioetica (i quattro principi presentati sopra) sia dal discorso sui diritti umani, che include, tra l'altro, il diritto alla libertà e alla dignità uguali davanti alla legge, la protezione dei diritti civili, politici e sociali, il riconoscimento universale della personalità giuridica e il diritto alla partecipazione libera e senza ostacoli alla vita della comunità (Leslie, 2019).

I quattro principi bioetici modificati con l'esplicabilità sono tradotti per l'IA in (Floridi et al., 2018) come segue:

1. Autonomia, intesa come il potere degli esseri umani di decidere se decidere, con il rischio di delegare troppo alle macchine.

2. Non maleficenza, intesa come prevenzione dei danni derivanti dall'intenzione degli esseri umani o dal comportamento imprevedibile delle macchine.

3. Beneficenza, intesa come promozione del benessere, preservazione della dignità e sostenibilità del pianeta.

4. Giustizia, intesa come prevenzione ed eliminazione delle discriminazioni ingiuste già esistenti, nonché dei nuovi danni, e garanzia di un'equa distribuzione dei benefici dell'IA.

5. Spiegabilità, definita come comprensione e responsabilità dei processi decisionali dell'IA.

Pertanto, l'etica dell'IA ha anche convergito su una serie di principi basati sui quattro principi classici dell'etica medica, nonché su altri approcci, riassunti in (Christoforaki &

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

Beyan, 2022). Tuttavia, come osservato in (Mittelstadt, 2019), rispetto alla medicina, lo sviluppo dell'IA manca di: (1) obiettivi comuni e doveri fiduciari, (2) storia e norme professionali, (3) metodi collaudati per tradurre i principi in pratica e (4) solidi meccanismi di responsabilità legale e professionale; ciò compromette il successo dell'approccio basato sui principi. Naturalmente, esiste anche un panorama normativo complesso che disciplina lo sviluppo e l'uso dell'IA nell'UE, comprese le leggi contro la discriminazione, un argomento che tuttavia esula dall'ambito della presente relazione.

Per quanto riguarda i diritti umani, secondo una relazione del 2018 finanziata dal Consiglio d'Europa (Comitato di esperti sugli intermediari Internet (MSI-NET), 2018), i diritti umani particolarmente interessati dagli algoritmi e dalle tecniche di trattamento automatizzato dei dati includono:

- Libera sperimentazione e giusto processo
- Privacy e protezione dei dati
- Libertà di espressione
- Ricorso effettivo
- Libertà di riunione e di associazione
- Divieto di discriminazione
- Diritti sociali e accesso ai servizi pubblici
- Diritto a elezioni libere

Gli algoritmi distorti sono citati esplicitamente come possibili fattori di discriminazione nei confronti di gruppi sociali in base all'età, all'orientamento sessuale, alla razza, al genere o alla posizione socioeconomica (Comitato di esperti sugli intermediari Internet

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

(MSI-NET), 2018, pag. 27). Inoltre, la Convenzione quadro del Consiglio d'Europa sull'intelligenza artificiale e i diritti umani, la democrazia e lo Stato di diritto menziona specificamente che gli Stati membri «adottano o mantengono misure volte a garantire che le attività nell'ambito del ciclo di vita dei sistemi di intelligenza artificiale rispettino l'uguaglianza, compresa l'uguaglianza di genere, e il divieto di discriminazione, come previsto dal diritto internazionale e nazionale applicabile» (Convenzione quadro del Consiglio d'Europa sull'intelligenza artificiale e i diritti umani, la democrazia e lo Stato di diritto, 2024, pag. 4).

Le organizzazioni della società civile (CSO), in quanto parti interessate nell'ecosistema sanitario (Vayena et al., 2018), possono svolgere un ruolo significativo nell'identificare e affrontare i pregiudizi dell'IA e la governance dell'IA in generale attraverso la promozione di uno sviluppo etico dell'IA, la responsabilizzazione delle parti interessate, l'educazione del pubblico, la rappresentanza delle comunità emarginate, la definizione di quadri politici e normativi e la promozione della collaborazione tra governi, aziende tecnologiche e pubblico (Korir, 2024).

All'interno di questo quadro teorico, sono state sviluppate esplicitamente una serie di soluzioni tecniche per affrontare i pregiudizi. Nella sezione seguente, presentiamo una classificazione dei pregiudizi indotti dall'IA che è servita come base per il nostro modello di mappatura, concentrandoci sul loro impatto sulla discriminazione di genere e razziale. I pregiudizi mentali umani (Hofmann, 2023), ad esempio i pregiudizi cognitivi, come il pregiudizio di conferma o di disponibilità, sebbene abbiano un forte impatto in medicina, sono considerati fuori dall'ambito del presente progetto.

Pregiudizi nei sistemi di IA

Il pregiudizio nei sistemi informatici è definito in (Friedman & Nissenbaum, 1996, p. 332) come un termine "[che si riferisce] a sistemi informatici che discriminano sistematicamente e ingiustamente determinati individui o gruppi di individui a favore

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

di altri. Un sistema discrimina ingiustamente se nega un'opportunità o un bene o se assegna un risultato indesiderabile a un individuo o a un gruppo di individui per motivi irragionevoli o inappropriati."

Secondo (Friedman & Nissenbaum, 1996), il pregiudizio nei sistemi informatici può essere suddiviso in tre categorie: pregiudizio preesistente, pregiudizio tecnico e pregiudizio emergente. Nella sottosezione seguente, esaminiamo ciascun tipo di pregiudizio e lo illustriamo con casi di studio, così come manifestato nella letteratura scientifica.

Pregiudizi preesistenti

I pregiudizi preesistenti derivano da pregiudizi nelle istituzioni sociali, nelle pratiche e negli atteggiamenti che già esistono e sono indipendenti e solitamente presenti prima della creazione del sistema. Questo tipo di pregiudizio viene incorporato nel sistema in modo consapevole o inconsapevole, a volte anche quando i creatori del sistema cercano di evitarlo.

Caso di studio: diagnosi delle malattie cardiovascolari nelle donne

Le malattie cardiovascolari (CVD) sono state comunemente percepite come "malattie maschili" e questo ha contribuito a una sottodiagnosi e a un trattamento insufficiente delle donne. Come dimostrato in (Al Hamid et al., 2024), una revisione sistematica sulla questione, le CVD erano meno segnalate tra le donne che mostravano sintomi più lievi rispetto agli uomini o che avevano sintomi erroneamente diagnosticati come sintomi gastrointestinali o legati all'ansia; pertanto, alle donne venivano offerti meno test diagnostici, farmaci e venivano indirizzate meno spesso a cardiologi e/o ricoveri ospedalieri. Inoltre, se ricoverate in ospedale, le donne erano meno propense a ricevere un intervento coronarico. Di conseguenza, i fattori di rischio delle donne erano sottovalutati dai medici, in particolare da quelli di sesso maschile. Dato che le donne continuano a essere sottorappresentate nel campo della cardiologia (Fatunde et

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

al., 2025), si può concludere che le donne sono meno propense a ricevere un'assistenza sanitaria adeguata a causa di pregiudizi già esistenti.

I sistemi di IA vengono addestrati utilizzando i dati raccolti dalle pratiche esistenti, quindi un sistema di diagnosi delle malattie cardiovascolari basato sull'IA incorporerà questo pregiudizio, creando una discriminazione nei confronti delle donne, indipendentemente dalle scelte effettuate durante l'implementazione tecnica.

Pregiudizio tecnico

Il pregiudizio tecnico deriva da vincoli o considerazioni di natura tecnica, in particolare quando i creatori di sistemi tentano di rendere i costrutti umani adatti ai computer, ad esempio quantificando il qualitativo, discretizzando il continuo o formalizzando il non formale. Inoltre, la decontestualizzazione degli algoritmi dagli ambienti in cui operano potrebbe impedire loro di trattare tutti i gruppi in modo equo in tutte le condizioni significative.

Caso di studio: Accuratezza predittiva dei modelli di previsione del rischio di ictus nelle popolazioni di razza bianca e nera

(Hong et al., 2023) hanno condotto uno studio retrospettivo sull'accuratezza predittiva del rischio di ictus confrontando i modelli di previsione del rischio specifici per l'ictus esistenti e le nuove tecniche di apprendimento automatico che coinvolgono, tra gli altri criteri, la razza dei pazienti. Tutti gli algoritmi hanno mostrato una discriminazione peggiore nei soggetti di razza nera rispetto a quelli di razza bianca. Secondo gli autori, questa situazione può essere attribuita a fattori di rischio non rilevati nei dati, come il tipo di assicurazione, le barriere linguistiche e altri fattori derivanti dal diverso accesso ai servizi sanitari, ovvero i dati sono decontestualizzati dall'ambiente socioeconomico in cui sono stati prodotti. Allo stesso tempo, tutti i fattori di rischio sopra menzionati

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

sono costrutti difficili da rappresentare in una forma accessibile ai computer. A tutto ciò si aggiunge il fatto che gli algoritmi di IA all'avanguardia sono per loro natura opachi riguardo alle caratteristiche che selezionano per ottenere un'elevata precisione (Knight, 2017), rendendo così anche i loro creatori incapaci di spiegare come funzionano e quindi di controllare se uno qualsiasi dei fattori socioeconomici sopra menzionati sia realmente preso in considerazione nel funzionamento interno del sistema di IA.

Pregiudizio emergente

Il bias emergente si manifesta in un contesto di utilizzo con utenti reali, in genere dopo il completamento della progettazione, come risultato di un cambiamento delle conoscenze sociali che non può essere, o non è, incorporato nella progettazione del sistema, o di una popolazione con conoscenze o valori culturali diversi da quelli ipotizzati nella progettazione.

Caso di studio: cambiamenti nei set di dati

Uno spostamento del set di dati è una discrepanza tra le distribuzioni dei set di dati di addestramento e di test durante lo sviluppo dell'algoritmo e può portare a prestazioni disparate a livello di sottogruppo (Chen et al., 2023).

Nel rilevamento del cancro della pelle, ad esempio, molti set di dati di imaging utilizzati per addestrare gli algoritmi di IA a rilevare il cancro della pelle provengono da paesi con popolazioni dalla pelle chiara (Guo et al., 2021), sottorappresentando così alcuni gruppi demografici. Gli algoritmi di IA addestrati con questi set di dati hanno prestazioni inferiori quando vengono applicati in paesi con una popolazione più diversificata, discriminando le persone dalla pelle scura. I set di dati sono difficili e costosi da raccogliere, annotare e convalidare, rendendo necessario per i sistemi di IA sviluppati nei paesi a basso e medio reddito affidarsi a set di dati disponibili pubblicamente che potrebbero non riflettere la distribuzione della loro popolazione,

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

con conseguente disallineamento tra la popolazione di origine e quella di destinazione. Lo stesso può verificarsi anche nei paesi ad alto reddito, ad esempio a causa dei cambiamenti demografici dovuti all'aumento dell'immigrazione o alle variazioni nella razza auto-dichiarata. Come osservato in (Chen et al., 2023), "poiché è ormai accettato che la razza sia un costrutto sociale e che vi sia una maggiore variabilità genetica all'interno di una particolare razza rispetto a quella tra le razze" [...] "la comunità medica ha iniziato a rendersi conto che le tassonomie del passato non rappresentano adeguatamente i gruppi di persone che pretendono di rappresentare" e "possono oscurare la cultura, la storia, lo status socioeconomico e altri fattori di confusione dell'equità".

Tipi di pregiudizi specifici della pipeline ML/AL

Sebbene quanto sopra sia valido per tutti i sistemi informatici, le applicazioni di IA hanno requisiti più specifici, quindi avevamo bisogno di una tassonomia più dettagliata. Di conseguenza, abbiamo deciso di seguire la classificazione dei pregiudizi presentata in (Suresh & Guttag, 2020), poiché identifica i tipi di pregiudizi in ogni fase della pipeline ML/AI, come illustrato nella Figura 1.

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

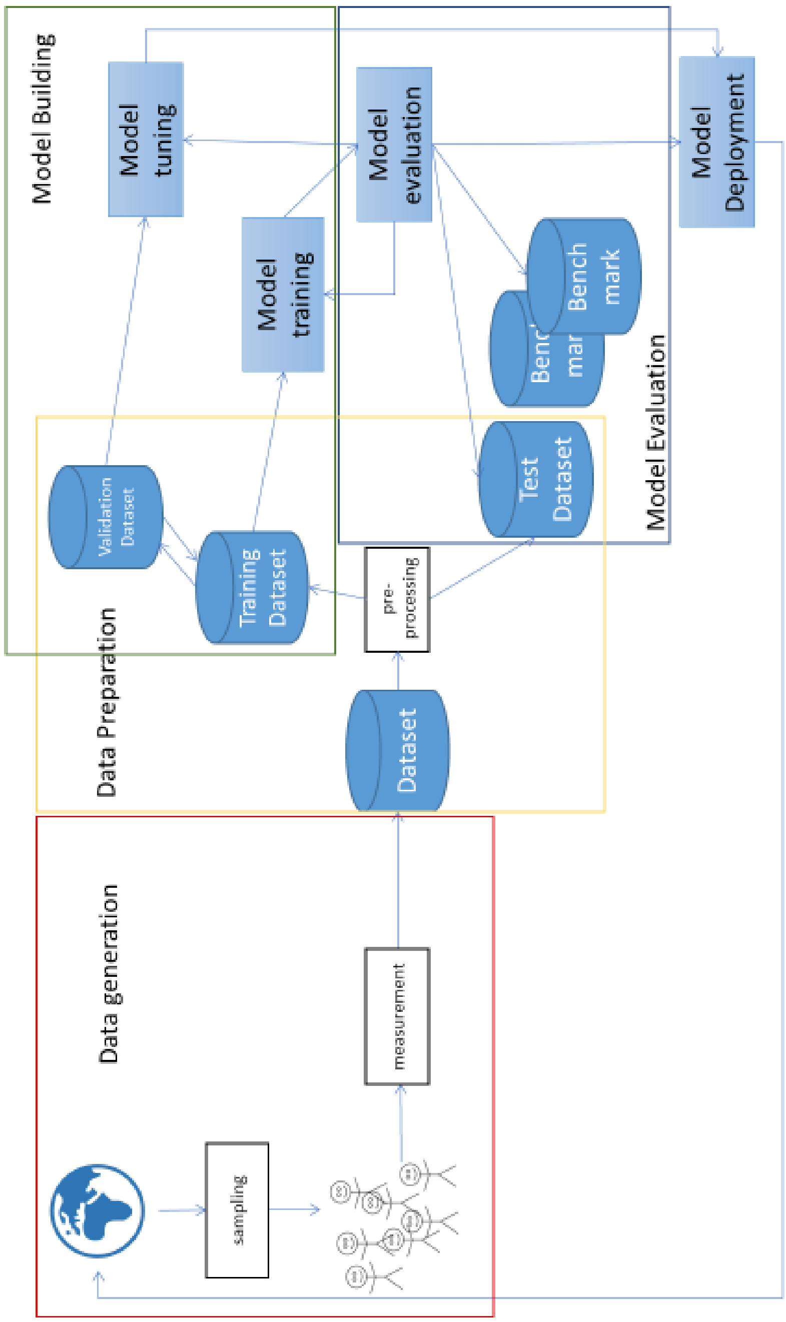

Figura 1 Pipeline ML/AI. Immagine adattata da (Suresh & Guttag, 2020)

Una tipica pipeline ML/AI può essere descritta come segue:

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

- **Generazione dei dati**. La creazione di un sistema ML/AI inizia con la generazione dei dati. Ciò comporta innanzitutto la raccolta e la preparazione dei dati per compilare un set di dati per il sistema AI. I dati esistenti nel mondo devono essere raccolti identificando un campione della popolazione target. La fase successiva consiste nel definire e misurare le caratteristiche rilevanti per l'applicazione da implementare e/o annotare i dati con etichette appropriate. Si tratta di un processo costoso e lungo, quindi il più delle volte i professionisti dell'AI utilizzano set di dati esistenti (pubblici o acquistati).

- **Preparazione dei dati**. In questa fase, il set di dati viene suddiviso in tre insiemi, ovvero: il *set di dati di addestramento*, ovvero il set di dati effettivo utilizzato per addestrare il modello; il *set di dati di convalida*, un campione di dati utilizzato per fornire una valutazione dell'adeguatezza di un modello al set di dati di addestramento durante la regolazione degli iperparametri del modello (parametri del modello che non possono essere appresi dai dati, ad esempio il numero di livelli e neuroni in un modello di rete neurale). In questa fase, potrebbe essere necessario pre-elaborare i dati (ad esempio, pulirli, normalizzarli); e il *set di dati di test*, la parte dei dati utilizzata per valutare il modello finale fornendo uno standard di riferimento una volta che il modello è completamente addestrato.

- **Costruzione del modello**. In questa fase, il modello viene addestrato sui dati di addestramento e messo a punto regolando gli iperparametri sul set di dati di convalida.

- **Valutazione del modello**. Il modello addestrato viene valutato utilizzando il set di dati di test e, talvolta, set di dati di benchmark, che sono set di dati compilati in modo indipendente utilizzati per dimostrare la robustezza del modello e/o consentire il confronto con altri metodi.

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

- **Implementazione** del **modello**. Applicazione del modello in un contesto reale. Ciò può portare a modifiche a seconda dei risultati e può anche creare un ciclo di feedback all'inizio della pipeline.

Tenendo conto delle fasi della pipeline ML/AI sopra descritte, adottiamo la classificazione dei pregiudizi di (Suresh & Guttag, 2020). Nello specifico, essi identificano le seguenti categorie di pregiudizi: pregiudizi storici, di rappresentazione, di misurazione, di aggregazione, di apprendimento, di valutazione e di implementazione. Nelle sottosezioni seguenti, definiamo i pregiudizi sopra elencati e proponiamo casi di studio tratti dalle fonti raccolte per il progetto.

Pregiudizio storico

Il bias storico corrisponde al bias preesistente definito da (Friedman & Nissenbaum, 1996), che incorpora pregiudizi e stereotipi già esistenti nei dati. Un esempio è riportato in (Calderone, 1990), che esamina se la frequenza della somministrazione di farmaci antidolorifici e sedativi ai pazienti sottoposti a intervento di bypass coronarico (CABG) postoperatorio differisca in base al sesso e all'età dei pazienti. Il risultato ha rivelato che ai pazienti di sesso maschile e ai pazienti di età pari o inferiore a 61 anni venivano somministrati farmaci antidolorifici con una frequenza significativamente maggiore rispetto alle pazienti di sesso femminile e ai pazienti di età pari o superiore a 62 anni, ai quali venivano invece somministrati farmaci sedativi con una frequenza significativamente maggiore. Il caso di studio sull'accuratezza predittiva dei modelli di previsione del rischio di ictus nelle popolazioni di razza nera e bianca lo dimostra nella sottosezione "Pregiudizi preesistenti"; tuttavia, presenteremo un altro caso di studio che mostra il pregiudizio storico relativo all'uso dell'IA nella salute mentale.

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

Caso di studio: l'intelligenza artificiale nella salute mentale e i pregiudizi dei modelli basati sul linguaggio

(Straw & Callison-Burch, 2020) presentano una revisione sistematica della letteratura sull'uso della NLP nella salute mentale, con l'obiettivo di identificare come questi pregiudizi possano ampliare le disuguaglianze sanitarie. I modelli di IA che utilizzano la NLP per profilare la salute mentale raccolgono grandi set di dati di linguaggio espressivo, tipicamente acquisiti dai social media, dai forum online, dai blog e dalle chat room. Tuttavia, questi dati sono già influenzati dal background personale e dal contesto sociale di un individuo.

In particolare, per quanto riguarda il genere e il linguaggio, esiste un'ampia bibliografia (sulla lingua inglese) riassunta in (Pennebaker et al., 2003), che rivela differenze nell'uso delle parole da parte delle donne e degli uomini. Ad esempio, le donne usano un linguaggio meno assertivo, che si manifesta in una maggiore cortesia, un minor uso di parolacce, un maggior uso di intensificatori (ad esempio, davvero, così) e un maggior uso di espressioni attenuanti (cioè qualificatori o parole incerte come tipo, forse o forse). Gli uomini, invece, sono stati descritti come direttivi, precisi e anche meno emotivi nell'uso del linguaggio, caratterizzato da riferimenti alla quantità, aggettivi giudicanti (ad esempio, buono, stupido), frasi ellittiche ("Bella foto") e riferimenti alla prima persona. Come osservano gli autori, queste differenze sono coerenti con un quadro sociologico delle differenze di genere, ma possono anche essere attribuite a spiegazioni alternative, come il maggiore impegno sociale delle donne.

Per quanto riguarda la salute mentale, uomini e donne scrivono biglietti di addio che esprimono il disagio suicida in modo diverso; le donne interiorizzano le emozioni negative, mentre gli uomini esprimono una rabbia crescente (Straw & Callison-Burch, 2020). Un sistema di intelligenza artificiale che individua i problemi di salute mentale per un genere può essere inadeguato per un altro (e questo considerando il genere in un contesto binario, che esclude gran parte della popolazione).

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

Pregiudizio di rappresentazione

Il bias di rappresentazione si verifica quando il campione di sviluppo sottorappresenta una parte della popolazione durante la fase di raccolta dei dati. Ciò può verificarsi nei seguenti modi: quando si definisce la popolazione target, se questa non riflette la popolazione di utilizzo; quando si definisce la popolazione target, se questa contiene gruppi sottorappresentati; quando si effettua il campionamento dalla popolazione target, se il metodo di campionamento è limitato o non uniforme. Il bias di rappresentazione comporta una scarsa generalizzazione per un sottoinsieme della popolazione di utenti. Un esempio tipico di bias di rappresentazione riguarda la diagnosi del cancro della pelle, poiché molti set di dati di imaging sottorappresentano alcuni gruppi demografici, causando l'addestramento dei modelli di apprendimento automatico su immagini di individui prevalentemente di carnagione chiara (Guo et al., 2021). Tenendo conto delle malattie target del progetto AEQUITAS, presentiamo un caso di studio sul bias di rappresentazione relativo alla razza nel diabete di tipo 2.

Caso di studio: valutazione del bias razziale negli algoritmi di previsione del rischio di diabete di tipo 2

Secondo (Cronjé et al., 2023), per quanto riguarda la popolazione statunitense, nonostante il loro rischio relativamente più basso, i gruppi bianchi non ispanici rimangono sovrarappresentati nella letteratura sulla previsione del rischio di diabete. In una diversa revisione sull'equità etnico-razziale nell'intelligenza artificiale per la gestione del diabete, negli articoli esaminati che riportavano la razza, la distribuzione media era del 69,5% di bianchi, del 17,1% di neri e del 3,7% di asiatici, mentre solo 2 articoli riportavano l'inclusione di partecipanti nativi americani (Pham et al., 2021).

È ben documentato che le disparità nei risultati del diabete sono in gran parte determinate da fattori sociali complessi e interconnessi che influenzano la salute, tra cui l'accesso a cibo sano, assistenza sanitaria di qualità, copertura assicurativa, barriere educative e tassi differenziali di adozione della tecnologia. Questi risultati includono

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

tassi più elevati di complicanze e un peggior controllo glicemico tra le popolazioni minoritarie e a basso reddito (Alipour & Alipour, 2025).

Di conseguenza, un sistema di IA addestrato su set di dati esistenti generalizzerebbe in modo inadeguato, portando a modelli predittivi distorti che potrebbero favorire individui di determinati gruppi razziali, ad esempio nelle azioni preventive.

Distorsione di misurazione

La distorsione di misurazione si verifica quando si scelgono, raccolgono o calcolano caratteristiche ed etichette da utilizzare in un problema di previsione, in particolare quando si utilizza un proxy (un'approssimazione di un costrutto che non è direttamente codificato o osservabile). Un esempio si trova in uno studio di Obermeyer et al. (2019), in cui i costi sanitari sono stati utilizzati come proxy per prevedere e classificare quali pazienti avrebbero tratto il massimo beneficio da cure extra, con conseguente discriminazione razziale. Tuttavia, i costi sanitari sono un proxy inadeguato per le esigenze sanitarie perché i pazienti di colore, che devono affrontare livelli di povertà sproporzionati, spesso spendono meno per l'assistenza sanitaria rispetto ai bianchi. A causa di questo bias, l'algoritmo ha erroneamente concluso che i pazienti di colore fossero più sani dei pazienti bianchi con lo stesso grado di malattia, classificandoli quindi come pazienti con priorità inferiore nell'accesso ai servizi sanitari.

Altre fonti di distorsione della misurazione possono verificarsi quando il metodo di misurazione varia da un gruppo all'altro, ad esempio quando due gruppi sono monitorati per lo stesso comportamento, ma uno dei due è monitorato in modo più rigoroso o frequente rispetto all'altro. Allo stesso modo, l'accuratezza della misurazione può variare da un gruppo all'altro, il che nelle applicazioni mediche può portare a tassi sistematicamente più elevati di diagnosi errate o sottodiagnosi in determinati gruppi. Ad esempio, i medici tendono a sottovalutare il dolore dei pazienti neri rispetto a quelli non neri a causa di false credenze sulle differenze biologiche tra

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

neri e bianchi, con la conseguenza che i pazienti neri hanno meno probabilità di ricevere farmaci antidolorifici e, se somministrati, ne ricevono quantità inferiori (Hoffman et al., 2016).

Caso di studio: differenze razziali ed etniche nell'associazione tra glucosio medio ed emoglobina A1c

Il test A1C misura la quantità media di glucosio (zucchero) nel sangue e viene utilizzato per rilevare il prediabete o aiutare a diagnosticare il diabete di tipo 2. Tuttavia, l'A1C è solo una misura indiretta e non è causalmente collegata agli esiti di salute, poiché esistono numerosi modi in cui la relazione tra le misure dirette della glicemia (la concentrazione di glucosio nel sangue) e l'A1C può essere direttamente alterata. Esiste anche una variazione sostanziale nella relazione glicemia-A1C tra gli individui e persino all'interno degli stessi individui nel corso del tempo. Inoltre, alcuni studi hanno riportato livelli significativamente più elevati di emoglobina A1c (A1C) nei pazienti afroamericani rispetto ai pazienti bianchi con lo stesso glucosio medio (Karter et al., 2023).

Se un sistema di IA progettato per diagnosticare il diabete viene addestrato a utilizzare i risultati del test A1C come proxy della glicemia senza tenere conto di altri fattori, come la razza del paziente, ciò può portare a diagnosi premature di diabete e trattamenti inappropriati, con conseguente qualità dell'assistenza sanitaria distorta e disuguaglianze sanitarie. Tuttavia, come osservato in (Alipour & Alipour, 2025), una revisione sistematica dei pregiudizi che potrebbero influenzare l'equità dei modelli di IA/ML nel diabete (incluso il pregiudizio di misurazione), mentre gli studi esaminati menzionano esplicitamente che il pregiudizio di misurazione può propagarsi attraverso i modelli di IA se non corretto, nessuno di essi ha tenuto conto di tali pregiudizi durante lo sviluppo del modello, li ha esplicitamente mitigati o ha segnalato la correzione delle differenze nell'accuratezza della misurazione.

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

Pregiudizio di aggregazione

Il bias di aggregazione si verifica quando si utilizza un modello unico per un set di dati che include gruppi diversi di persone o cose.

Possiamo considerare l'esempio della mappatura dei dati di input (ad esempio, il reddito di una persona) su etichette che li descrivono (ad esempio, basso, medio, alto) che si presume siano coerenti tra i sottoinsiemi dei dati. In realtà, il background o la cultura di una persona possono cambiare il significato effettivo di quei numeri. Ad esempio, un reddito "alto" in una piccola città rurale o in un paese a basso o medio reddito potrebbe avere un significato molto diverso rispetto a quello che ha in una grande città o in un paese ad alto reddito.

Caso di studio: strumenti sanitari digitali per il monitoraggio passivo della depressione

L'uso di strumenti digitali per misurare le variabili fisiologiche e comportamentali per il monitoraggio passivo della depressione è affrontato da (De Angel et al., 2022), una revisione sistematica sull'argomento. Gli articoli esaminati hanno analizzato le associazioni tra la depressione e i dati comportamentali oggettivi ottenuti dai sensori degli smartphone e dei dispositivi indossabili. Questi dati sono stati mappati in caratteristiche utilizzate dai modelli di IA per fare previsioni, corrispondenti al sonno, all'attività fisica, al ritmo circadiano, alla socievolezza, alla posizione e all'uso del telefono.

Tuttavia, gli autori sottolineano l'eterogeneità che deriva dalla diversità dei metodi utilizzati per creare queste caratteristiche. Ad esempio, la caratteristica "qualità del sonno" può essere definita misurando il numero di risvegli, il numero totale di minuti di veglia o la proporzione tra veglia e sonno in una sessione di sonno, mentre dobbiamo anche prendere in considerazione le differenze nel modo in cui i sensori dei diversi dispositivi descrivono un evento come "sonno". Poiché tutte le differenziazioni

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

di cui sopra non vengono prese in considerazione e vengono raggruppate collettivamente come "qualità del sonno", e poiché un set di dati potrebbe provenire da persone o gruppi con background, culture o norme diverse, questa caratteristica può avere un significato diverso per ciascuno di questi gruppi o individui.

L'aggregazione di tali dati in un'unica caratteristica può portare a un sistema che non si adatta a nessun gruppo o che privilegia la popolazione dominante se esiste anche un pregiudizio di rappresentanza. Ad esempio, è dimostrato che esistono differenze di genere nel sonno tra uomini e donne, mentre queste ultime sono spesso sottorappresentate nella ricerca sul sonno. Inoltre, altri fattori che di solito non vengono presi in considerazione per i modelli e i disturbi del sonno sono la mancata distinzione tra genere come costrutto sociale e sesso biologico e la mancata considerazione delle identità intersezionali definite da età, razza e classe socioeconomica (Lok et al., 2024).

Pregiudizio di apprendimento

Il bias di apprendimento si verifica quando le scelte di modellizzazione amplificano le disparità di performance tra i diversi esempi presenti nei dati. Un esempio riguarda la privacy differenziale, un meccanismo utilizzato nei sistemi di IA che garantisce che, esaminando l'output di un sistema, non sia possibile determinare se i dati di un individuo specifico siano stati inclusi nel set di dati originale. La privacy differenziale viene utilizzata nei set di dati sanitari per proteggere le informazioni sensibili dei pazienti, ad esempio nel caso di malattie rare, dove il caso di ogni paziente è più o meno unico in un'area limitata coperta da un ospedale, quindi anche se i dati sono resi anonimi, non è molto difficile dedurre l'identità della persona. Tuttavia, è stato dimostrato che la privacy differenziale riduce l'influenza dei dati sottorappresentati sul modello; pertanto, se il sistema di IA è inizialmente distorto, l'applicazione di una misura di rafforzamento della privacy aggrava ulteriormente tale distorsione (Bagdasaryan & Shmatikov, 2019).

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

Caso di studio: privacy differenziale e disparità sanitarie

Nel settembre 2018, l'Ufficio censimento degli Stati Uniti ha annunciato che avrebbe implementato la privacy differenziale sui prodotti di dati derivati dai dati del censimento del 2020. Tuttavia, (Santos-Lozada et al., 2020) hanno studiato il modo in cui l'implementazione della privacy differenziale può alterare la conoscenza delle disparità sanitarie in termini di mortalità, in particolare per le minoranze razziali o etniche in aree piccole e contesti meno urbani. I loro risultati hanno suggerito che la privacy differenziale influenzerà in modo più marcato le stime del tasso di mortalità per i neri non ispanici e gli ispanici rispetto alle stime per i bianchi non ispanici.

Questi risultati sono stati confermati da (Kurz et al., 2022), che dimostrano che l'applicazione della privacy differenziale agli stessi dati può portare a una rappresentazione errata dei tassi di partecipazione al programma Medicaid tra i gruppi razziali ed etnici già emarginati. Nello specifico, questi tassi per alcune combinazioni di contea, razza ed etnia differivano tra i risultati dei dati con privacy differenziale e i dati originali, superando talvolta il 10%. Inoltre, gli individui bianchi non ispanici erano l'unico sottogruppo etnico e razziale per il quale l'algoritmo di privacy differenziale ha catturato accuratamente i tassi di partecipazione al programma Medicaid. Questo risultato può avere importanti implicazioni per la politica sanitaria, poiché i dati del censimento vengono utilizzati per pianificare i programmi governativi, allocare le risorse e valutare e monitorare le politiche.

Distorsione di valutazione

Il bias di valutazione si verifica quando i dati di riferimento utilizzati per un compito particolare non rappresentano la popolazione di utilizzo. I benchmark sono set di dati standardizzati utilizzati per misurare la qualità di un modello, consentendo il confronto quantitativo dei modelli. Di conseguenza, esiste il rischio di incoraggiare lo sviluppo e l'implementazione di modelli che funzionano bene solo sul sottoinsieme dei dati rappresentati nel benchmark. Pertanto, se il benchmark è soggetto a bias storici,

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

rappresentativi o di misurazione, può verificarsi una discriminazione nei confronti di sottogruppi o individui vulnerabili.

Nel settore sanitario, le ragioni della sottorappresentazione di popolazioni specifiche nei set di dati possono essere dovute all'assenza di individui o gruppi dai set di dati (ad esempio, le donne in gravidanza, a causa di vincoli etici) o alla categorizzazione errata o inappropriata delle persone in gruppi (ad esempio, categorie di "etnia mista" o "altro"). Le cause alla base di questo fenomeno possono includere ragioni sociali e tecniche o legali/etiche, quali barriere strutturali all'accesso all'assistenza sanitaria, ostacoli tecnici alla raccolta o alla digitalizzazione dei dati sanitari rilevanti, limitazioni individuali e strutturali relative al consenso alla condivisione dei dati e restrizioni legali o etiche alla condivisione dei dati che ne impediscono l'accessibilità, tra le altre (Arora et al., 2023). Il risultato è che i sistemi di IA calibrati su tali parametri di riferimento possono avere prestazioni inferiori quando applicati a individui appartenenti a un gruppo sottorappresentato. Tuttavia, è importante notare che la validità dei parametri di riferimento è una questione più generica e non si limita al pregiudizio (Brooks, 2025).

Caso di studio: set di dati di immagini della pelle

I set di dati di immagini della pelle sottorappresentano alcuni gruppi demografici, poiché la maggior parte delle immagini in questi set di dati proviene da popolazioni del Nord America o dell'Europa e raffigura prevalentemente individui con pelle chiara (Guo et al., 2021). A causa dell'alto costo e della difficoltà di costruire questi set di dati, oltre che per l'addestramento dei modelli, essi possono essere utilizzati anche come benchmark.

Il caso di studio che illustra il pregiudizio emergente, ovvero i set di dati di immagini di tumori della pelle utilizzati per addestrare modelli di previsione, è un esempio di benchmark inappropriato quando la popolazione di utenti proviene da gruppi

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

sottorappresentati (Guo et al., 2021). Un caso simile, sebbene non correlato all'IA, mostra la generalità del problema, che riguardava i pulsossimetri (dispositivi che misurano la saturazione di ossigeno nel sangue, utilizzati, ad esempio, in caso di infarto o insufficienza cardiaca), che hanno dimostrato di funzionare in modo più accurato sulla pelle chiara (Sjoding et al., 2020).

I pregiudizi relativi alla rappresentazione, alla misurazione, all'aggregazione, all'apprendimento e alla valutazione possono essere ricondotti al pregiudizio tecnico definito da (Friedman & Nissenbaum, 1996).

Pregiudizio di implementazione

Il bias di implementazione si verifica quando c'è una discrepanza tra il problema che un modello intende risolvere e il modo in cui viene effettivamente utilizzato, il che può causare danni, soprattutto se combinato con bias cognitivi come il bias di conferma e il bias di automazione. Il bias di implementazione è lo stesso del bias emergente definito da (Friedman & Nissenbaum, 1996).

Caso di studio: Cambiamento di dominio

Il caso del cambiamento di dati è documentato nella sottosezione sui pregiudizi emergenti relativa alla diagnosi del cancro della pelle. Inoltre, possiamo definire il caso del cambiamento di dominio, che si verifica quando un sistema viene implementato, ha superato l'autorizzazione normativa ed è utilizzato nella pratica clinica, ma viene applicato a una coorte di pazienti diversa da quella per cui è stato addestrato. Ad esempio, un sistema può essere sviluppato per un ospedale in un paese ad alto reddito ed essere implementato in un paese a basso o medio reddito senza tenere conto di fattori quali le caratteristiche sociodemografiche dei pazienti o se questi ultimi hanno

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

lo stesso livello di rischio complessivo rispetto a quelli inclusi nei dati di addestramento (Vokinger et al., 2021).

Implicazioni politiche

Le prove mappate nel Deliverable D2.1 dimostrano che i pregiudizi di genere e razziali nell'IA biomedica non sono difetti tecnici accidentali o isolati, ma rischi sistemici che emergono durante l'intero ciclo di vita dei sistemi di IA utilizzati nell'assistenza sanitaria. Nelle malattie cardiovascolari, nella depressione e nel diabete, i pregiudizi derivano da set di dati clinici storicamente distorti, pratiche diagnostiche inique, variabili proxy che codificano le disuguaglianze strutturali e contesti di implementazione che distribuiscono in modo non uniforme sia i benefici che i danni. Questi risultati confermano che l'IA biomedica coinvolge direttamente molteplici diritti e principi tutelati dalla Carta dei diritti fondamentali dell'Unione europea, in particolare i precetti della dignità umana, dell'uguaglianza davanti alla legge e della non discriminazione, nonché il diritto all'integrità della persona, il diritto all'assistenza sanitaria, la protezione dei dati e il diritto a un ricorso effettivo.

In questo contesto, i quadri politici dell'UE e nazionali che disciplinano l'IA nell'assistenza sanitaria devono considerare la mitigazione dei pregiudizi non come un'aggiunta etica volontaria, ma come una componente vincolante di un'implementazione dell'IA legale e conforme ai diritti. Gli sforzi normativi europei e nazionali in materia di IA nell'assistenza sanitaria dovrebbero essere considerati nell'ambito del più ampio quadro dei diritti fondamentali che disciplina l'IA (cfr. Novossiolova, 2025; Novossiolova et al., 2025; Kasapi, 2025). La legge dell'UE sull'IA fornisce una struttura normativa necessaria classificando la maggior parte dell'IA biomedica come sistemi ad alto rischio, ma la sua efficacia nella pratica dipenderà dal modo in cui le garanzie dei diritti fondamentali saranno attuate nelle valutazioni di conformità, nel monitoraggio post-commercializzazione e negli appalti pubblici.

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

In primo luogo, le garanzie di un controllo umano significativo devono essere rafforzate e specificate per i sistemi di IA biomedica durante tutto il loro ciclo di vita. Gli strumenti clinici di IA utilizzati per la diagnosi, la stratificazione del rischio, lo screening o il supporto terapeutico non devono in alcun caso funzionare come decisori autonomi de facto. La supervisione umana deve includere non solo la possibilità di override da parte degli operatori sanitari, ma anche una chiara responsabilità istituzionale per la comprensione dei limiti del sistema, dei rischi di bias noti e dei divari di performance dei sottogruppi. In linea con la tutela della dignità e dell'integrità umana sancita dalla Carta, gli operatori sanitari dovrebbero essere formati e sostenuti istituzionalmente per interrogare criticamente i risultati dell'IA piuttosto che deferire ad essi. Ciò richiede l'integrazione dell'alfabetizzazione all'IA, della consapevolezza dei bias e della formazione sui diritti fondamentali nell'educazione medica e nello sviluppo professionale continuo.

Gli obblighi di trasparenza dovrebbero essere interpretati in modo estensivo nei contesti sanitari. I pazienti e gli utenti dei servizi sanitari devono essere informati ogni volta che i sistemi di IA vengono utilizzati nel processo decisionale clinico che li riguarda, compresi lo screening, la definizione delle priorità o la valutazione del rischio. Quando i risultati generati dall'IA informano i servizi sanitari pubblici, tali risultati dovrebbero essere chiaramente identificabili come tali e accompagnati da spiegazioni accessibili sul loro ruolo, sui loro limiti e sui rischi di parzialità noti. Gli individui dovrebbero inoltre essere informati quando i loro dati personali vengono utilizzati per la formazione, il collaudo o l'apprendimento continuo dell'IA, in particolare quando sono coinvolti dati sanitari sensibili. Queste misure di trasparenza sono essenziali per sostenere i diritti sanciti dalla Carta in materia di protezione dei dati e di ricorso effettivo e per consentire agli individui di contestare in modo significativo le decisioni che potrebbero avere un impatto negativo su di loro.

In secondo luogo, la valutazione dell'impatto sui diritti fondamentali deve diventare un requisito sistematico e applicabile per i sistemi di IA biomedici, che vada oltre i

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

controlli pre-commercializzazione fino alla valutazione continua durante l'implementazione. Le prove empiriche riportate in D2.1 dimostrano che molti danni causati dai pregiudizi diventano visibili solo quando i sistemi di IA interagiscono con popolazioni reali e flussi di lavoro clinici, in particolare attraverso effetti intersezionali che coinvolgono genere, razza, età e status socioeconomico. Le valutazioni d'impatto basate sui diritti, come quelle ispirate alla metodologia HUDERIA del Consiglio d'Europa (Metodologia per la valutazione dei rischi e dell'impatto dei sistemi di intelligenza artificiale dal punto di vista dei diritti umani, della democrazia e dello Stato di diritto), dovrebbero quindi essere obbligatorie per l'IA medica ad alto rischio, esaminando esplicitamente le differenze di prestazioni e risultati tra i gruppi protetti. Tali valutazioni devono prevedere una partecipazione significativa delle parti interessate, comprese le organizzazioni della società civile, i rappresentanti dei pazienti e gli organismi per la parità, al fine di far emergere i danni che potrebbero essere invisibili da un punto di vista puramente tecnico o clinico.

Dovrebbero essere richiesti audit periodici dei sistemi di IA biomedica per verificare la continua conformità agli standard dei diritti fondamentali, con particolare attenzione alla deriva dei pregiudizi, ai cambiamenti nei set di dati e ai cambiamenti nell'uso clinico nel tempo. Laddove gli audit rivelino effetti discriminatori persistenti o non mitigabili, devono esserci percorsi legali e istituzionali chiari per limitare, sospendere o interrompere l'uso del sistema. Il diritto all'assistenza sanitaria non può giustificare il continuo impiego di strumenti di IA che svantaggiano sistematicamente determinati gruppi, anche se le metriche aggregate delle prestazioni appaiono favorevoli.

In terzo luogo, le autorità dell'UE e nazionali devono affrontare il rischio di uso improprio e danni secondari associati all'IA biomedica. Ciò include le vulnerabilità della sicurezza informatica che potrebbero compromettere l'integrità del sistema o consentire la manipolazione dolosa dei risultati clinici, nonché il riutilizzo dell'IA sanitaria per la sorveglianza, la profilazione o pratiche di esclusione. I sistemi di IA biomedica dovrebbero essere soggetti a regolari valutazioni di sicurezza e a solidi

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

obblighi di segnalazione degli incidenti, con chiari meccanismi di responsabilità nei casi in cui sistemi distorti o compromessi portino a violazioni dei diritti. I quadri di responsabilità dovrebbero garantire che la responsabilità non possa essere attribuita esclusivamente ai singoli medici quando i danni sono strutturalmente incorporati nella progettazione dell'IA o nelle decisioni di implementazione.

In quarto luogo, la promozione di pratiche etiche e responsabili deve essere integrata nell'intera catena del valore dell'IA biomedica. Gli sviluppatori dovrebbero essere tenuti ad affrontare in modo proattivo i rischi di parzialità attraverso la raccolta di dati rappresentativi, un'attenta selezione degli obiettivi e dei proxy, la convalida specifica per sottogruppi e la rendicontazione trasparente delle prestazioni tra i gruppi di genere e razziali. È importante sottolineare che le prove esaminate in D2.1 dimostrano che la "correttezza attraverso l'inconsapevolezza" e le strategie di eliminazione dei pregiudizi puramente tecniche sono spesso insufficienti in ambito sanitario. Le linee guida e gli standard normativi dovrebbero quindi andare oltre le metriche astratte di equità e richiedere agli sviluppatori di dimostrare risultati clinicamente significativi in termini di equità, valutati in relazione ai percorsi sanitari reali e ai modelli di accesso.

Le politiche di appalto pubblico e di finanziamento svolgono un ruolo cruciale nel plasmare gli incentivi per gli sviluppatori. Le autorità sanitarie e gli ospedali pubblici dovrebbero integrare i diritti fondamentali e i criteri di pregiudizio nelle decisioni di appalto per i sistemi di IA, favorendo soluzioni che dimostrino pratiche di mitigazione dei pregiudizi solide, trasparenti e verificate in modo indipendente. Gli strumenti di finanziamento dell'UE, compresi i futuri programmi di ricerca e innovazione, dovrebbero continuare a dare priorità ai progetti che combinano l'innovazione tecnica con la governance basata sui diritti, il coinvolgimento delle parti interessate e lo sviluppo di capacità, in linea con il modello AEQUITAS.

Infine, il rafforzamento della resilienza sociale nei confronti dell'IA biomedica discriminatoria richiede investimenti sostenuti nella sensibilizzazione dell'opinione

Finanziato dall'Unione Europea. Le opinioni e i pareri espressi sono tuttavia esclusivamente quelli dell'autore/degli autori e non riflettono necessariamente quelli della Commissione Europea-UE. Né l'Unione Europea né la Commissione Europea possono essere ritenute responsabili per essi.
Codice del progetto: 101191047 — 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Cofinanziato dall'Unione europea

AEQUITAS

pubblica, nel coinvolgimento della società civile e nella collaborazione intersettoriale. Gli individui devono essere messi in grado di comprendere i propri diritti nell'assistenza sanitaria mediata dall'IA e i meccanismi disponibili per proteggerli. Le organizzazioni della società civile, gli organismi per la parità e i gruppi di pazienti dovrebbero essere riconosciuti come attori essenziali nel monitoraggio degli impatti dell'IA, nel sostegno alle persone colpite e nell'informazione sullo sviluppo delle politiche. La cooperazione tra governi, operatori sanitari, ricercatori, industria e società civile è necessaria per garantire che i benefici dell'IA biomedica siano condivisi equamente e non rafforzino le disuguaglianze sanitarie esistenti.

Nel loro insieme, i risultati del Deliverable D2.1 supportano una chiara conclusione politica: l'IA biomedica può essere considerata affidabile e legittima nell'UE solo quando la sua progettazione, implementazione e governance sono saldamente ancorate alla protezione dei diritti fondamentali. La legge dell'UE sull'IA, interpretata alla luce della Carta dei diritti fondamentali dell'Unione europea e resa operativa attraverso meccanismi concreti di supervisione, valutazione d'impatto e responsabilità, offre un'opportunità fondamentale per garantire che l'innovazione nel settore sanitario promuova l'equità anziché riprodurre modelli storici di discriminazione.