### Introduzione

L'intelligenza artificiale (AI), definita come la teoria e lo sviluppo di sistemi informatici capaci di eseguire compiti che solitamente richiedono l'intelligenza umana, rappresenta una rivoluzione in atto destinata a trasformare radicalmente scienza e società. Approcci innovativi come il machine learning, il deep learning e le reti neurali artificiali stanno rimodellando l'elaborazione e l'analisi dei dati, mentre i sistemi autonomi e semi-autonomi trovano crescente applicazione in settori come sanità, trasporti e produzione industriale.

L'impatto trasformativo dell’AI e le sue implicazioni sociali hanno stimolato un vivace dibattito sui principi etici e sui valori che dovrebbero orientarne sviluppo e utilizzo. Tra le principali preoccupazioni emergono il rischio di perdita di posti di lavoro, l'uso malevolo della tecnologia, la difficoltà di attribuzione di responsabilità e la potenziale diffusione involontaria di bias, compromettendo l'equità. Questi temi sono ampiamente discussi nella letteratura scientifica recente e nei media.

Numerose organizzazioni nazionali e internazionali hanno istituito comitati di esperti dedicati all’AI, incaricati di redigere linee guida e documenti di politica. Tra questi, il Gruppo di Esperti di Alto Livello sull'Intelligenza Artificiale della Commissione Europea, il gruppo di esperti sull’AI dell'OCSE, il Consiglio Consultivo sull'Uso Etico dell'Intelligenza Artificiale e dei Dati di Singapore e il Comitato Selezionato sull'Intelligenza Artificiale della Camera dei Lord del Regno Unito. Questi comitati hanno prodotto o stanno elaborando rapporti e documenti di orientamento per indirizzare lo sviluppo e l'uso etico dell’AI. Anche il settore privato ha intrapreso iniziative simili; nel 2018, ad esempio, aziende come Google e SAP hanno pubblicato proprie linee guida e principi sull’AI. Associazioni professionali e organizzazioni non profit, tra cui l'Association for Computing Machinery (ACM), Access Now e Amnesty International, hanno emesso dichiarazioni e raccomandazioni. Questa proliferazione di iniziative di soft-law può essere interpretata come una risposta di governance alla rapida evoluzione della ricerca sull’AI, accompagnata da un significativo aumento della produzione scientifica e delle dimensioni del mercato.

Dall'analisi del contenuto di vari documenti di ricerca emerge un insieme di undici valori e principi etici ricorrenti. Questi, ordinati per frequenza di menzione nelle fonti, sono: trasparenza, giustizia ed equità, non maleficenza, responsabilità, privacy, beneficenza, libertà e autonomia, fiducia, dignità, sostenibilità e solidarietà. Sebbene nessun singolo principio etico sia risultato comune a tutti i documenti esaminati, si osserva una convergenza attorno a cinque principi principali: trasparenza, giustizia ed equità, non maleficenza, responsabilità e privacy, citati in oltre la metà delle fonti analizzate.

Con l'aumento della disponibilità di dati e la sofisticazione delle tecniche di machine learning (ML), sono aumentate le opportunità di migliorare sia la vita pubblica che privata. La società ha un controllo maggiore sugli esiti legati a chi le persone possono diventare, cosa possono fare, cosa possono ottenere e come possono interagire con il mondo. Tuttavia, le crescenti preoccupazioni riguardo alle sfide etiche poste dall'uso del ML e dell’AI in generale minacciano di arrestare l'avanzamento delle applicazioni benefiche, a meno che non vengano gestite correttamente. Bilanciare la tensione tra supportare l'innovazione e limitare i potenziali danni associati all’AI mal progettata è una sfida. Gli algoritmi di ML sono potenti costrutti socio-tecnici che sollevano preoccupazioni sia per le persone che per il codice.

Abilitare il doppio vantaggio del 'ML etico'—capitalizzando le opportunità e minimizzando i danni—richiede di affrontare domande difficili su progettazione, sviluppo, distribuzione, pratiche, usi e utenti, oltre che sui dati che alimentano l'intero ciclo di vita degli algoritmi. È possibile progettare una società arricchita dagli algoritmi in modo pro-etico, proteggendo i valori, i principi e l'etica che la società ritiene fondamentali. Questo è il messaggio che scienziati sociali, eticisti, filosofi, politici, tecnologi e società civile stanno consegnando in un appello collettivo per lo sviluppo di meccanismi di governance appropriati che permetteranno alla società di capitalizzare le opportunità garantendo nel contempo che i diritti umani siano rispettati e che le decisioni siano giuste ed etiche.

### Standard Intergovernativi sull’AI

Il 22 maggio 2019, l'OCSE ha annunciato che i suoi trentasei paesi membri, insieme a sei ulteriori (Argentina, Brasile, Colombia, Costa Rica, Perù e Romania), avevano formalmente adottato il primo standard intergovernativo sull’AI. Questo standard consiste in cinque principi complementari basati sui valori e cinque raccomandazioni implementabili per i decisori politici. Questi principi e raccomandazioni sono frutto di numerosi documenti pubblicati negli ultimi tre anni, con l'obiettivo di fungere da vincoli normativi sull'uso degli algoritmi nella società.

Una revisione di 84 documenti etici sull’AI ha rilevato che, sebbene nessun principio singolo fosse presente in tutti i documenti, i temi della trasparenza, giustizia ed equità, non maleficenza, responsabilità e privacy comparivano in oltre la metà di essi. Analogamente, una revisione sistematica della letteratura sulla tecnologia etica ha rivelato che i temi della privacy, sicurezza, autonomia, giustizia, dignità umana, controllo della tecnologia e bilanciamento dei poteri erano ricorrenti. Questo consenso offre una base condivisa su cui costruire, utile per comunicare aspettative e valutare risultati. Tuttavia, la disponibilità di questi principi "concordati" non garantisce ancora un cambiamento reale nella progettazione dei sistemi algoritmici.

### La Sfida di un ML Etico

Scienziati sociali e filosofi politici sono abituati a gestire la pluralità e soggettività che informa l'etica del ML. Rispondere a domande come cosa succede quando l'etica a livello individuale e di gruppo interagiscono e cosa significano termini come 'equità', 'responsabilità', 'trasparenza' e 'interpretabilità' richiede una visione contestuale della privacy e della trasparenza. La comunità degli sviluppatori ML, meno abituata a tali complessità, deve affrontare il trilemma morale-semantico, poiché gli strumenti esistenti sono spesso troppo rigidi, troppo flessibili o imprevedibili.

### Collaborazione Multidisciplinare per un'Etica dell’AI

Coinvolgere ricercatori multidisciplinari nello sviluppo di strumenti e metodologie pro-etici è essenziale. Un approccio multidisciplinare aiuterà la comunità etica ML a superare gli ostacoli riguardanti la complessità sociale, abbracciare l'incertezza e accettare che l’AI è costruita su ipotesi, il comportamento umano è complesso, gli algoritmi possono avere conseguenze ingiuste e le previsioni algoritmiche possono essere difficili da interpretare. Le collaborazioni risultanti saranno altamente benefiche per lo sviluppo di strumenti e metodologie di etica applicata per almeno tre ragioni: garantiranno la protezione del pluralismo dei valori, incoraggeranno un approccio più riflessivo all'etica applicata e mitigheranno il rischio di un'applicazione sporadica di strumenti etici durante lo sviluppo.

Questo approccio dipenderà dalla creazione di più strumenti e dall'accelerazione del loro sviluppo dai laboratori di ricerca agli ambienti di produzione. La società deve riunirsi in comunità multidisciplinari per promuovere lo sviluppo di conoscenza e comprensione comuni e un obiettivo condiviso per l'etica applicata dell’AI.

In definitiva, questo produrrà risultati migliori e incoraggerà la competizione per creare strumenti che colmino le lacune esistenti. La comunità di ricerca ML dovrebbe collaborare con un focus su:

1. lo sviluppo di un linguaggio comune;
2. la creazione di strumenti che garantiscano una partecipazione equa alla progettazione delle soluzioni algoritmiche;
3. la valutazione degli strumenti esistenti per identificarne efficacia e aree di miglioramento;
4. un impegno per la riproducibilità, l'apertura e la condivisione della conoscenza e delle soluzioni tecniche;
5. la creazione di esempi pratici di utilizzo degli strumenti etici;
6. la valutazione e la creazione di modelli di business pro-etici che bilancino i costi e i benefici dell'investimento in AI etica.

### Principi Etici nell’AI

#### Trasparenza

Questo principio è il più prevalente nella letteratura attuale. L'interpretazione della trasparenza varia significativamente in relazione alla sua giustificazione, dominio di applicazione e modalità di implementazione. Le fonti collegano la trasparenza a sforzi per aumentare la spiegabilità e l'interpretabilità, nonché ad atti di comunicazione e divulgazione. I principali domini di applicazione includono l'uso dei dati, l'interazione uomo-IA, le decisioni automatizzate e lo scopo dell'uso dei dati o dell'applicazione dei sistemi AI. La trasparenza è vista principalmente come un mezzo per minimizzare i danni e migliorare l’AI, anche se alcuni studi ne evidenziano i benefici legali o per promuovere la fiducia. Viene anche collegata a dialogo, partecipazione e principi democratici.

Per migliorare la trasparenza, molte fonti suggeriscono una maggiore divulgazione delle informazioni da parte degli sviluppatori e utilizzatori di sistemi AI. Le specifiche su cosa debba essere comunicato variano: uso dell’AI, codice sorgente, uso dei dati, base di evidenze per l'uso dell’AI, limitazioni, leggi, responsabilità, investimenti e potenziale impatto. Viene incoraggiata la fornitura di spiegazioni "in termini non tecnici" o audibili da parte degli esseri umani. Mentre audit e auditabilità sono principalmente proposti da uffici di protezione dei dati e organizzazioni non profit, è il settore privato a suggerire soluzioni tecniche. Altre misure si concentrano su supervisione, interazione e mediazione con stakeholder e pubblico, e facilitazione del whistleblowing.

#### Giustizia, Equità e Imparzialità

La giustizia è espressa principalmente in termini di equità e prevenzione, monitoraggio o mitigazione di bias indesiderati e discriminazione, sebbene quest'ultima sia meno menzionata dal settore privato. Alcune fonti interpretano la giustizia come rispetto della diversità, inclusione e uguaglianza, mentre altre richiedono la possibilità di appellarsi o contestare le decisioni, nonché il diritto al risarcimento e alla riparazione. È evidenziata anche l'importanza di un accesso equo ai benefici dell’AI. Gli emittenti del settore pubblico sottolineano l'impatto dell’AI sul mercato del lavoro e la necessità di affrontare questioni democratiche o sociali. La giustizia viene promossa attraverso soluzioni tecniche, trasparenza, test, monitoraggio, audit, sviluppo o rafforzamento dello stato di diritto e processi sistemici e partecipativi.

#### Non Maleficenza

I riferimenti alla non maleficenza sono più frequenti rispetto alla beneficenza e includono richiami alla sicurezza e alla protezione, affermando che l’AI non dovrebbe mai causare danni prevedibili o non intenzionali. I danni specifici da evitare includono discriminazione, violazione della privacy e danni fisici, con considerazioni dettagliate su cyberwarfare e hacking malevolo. Le linee guida per la prevenzione dei danni si concentrano su misure tecniche e strategie di governance, che vanno dalla ricerca sull’AI al design, sviluppo tecnologico e implementazione, fino a valutazioni continue e processi di audit. Viene suggerita la cooperazione interdisciplinare, la conformità con le leggi esistenti o nuove, e l'istituzione di pratiche di supervisione, con un'attenzione particolare alla valutazione dei rischi e alla chiara attribuzione delle responsabilità. Alcune fonti mettono in guardia contro gli usi multipli o doppi usi, e altre si oppongono esplicitamente all'applicazione militare dell’AI.

#### Responsabilità e Accountability

Nonostante i diffusi riferimenti a una "IA responsabile", la responsabilità e l'accountability sono raramente definite chiaramente. Alcune raccomandazioni includono l'agire con integrità e la chiarificazione delle responsabilità legali nei contratti o in anticipo. Altre fonti suggeriscono di concentrarsi sui processi che possono portare a comportamenti dannosi, piuttosto che su chi dovrebbe essere ritenuto responsabile dopo che i danni si sono verificati. La responsabilità viene spesso collegata a motivi e processi sottostanti, enfatizzando la necessità di interventi proattivi piuttosto che reattivi.

### Focus sulla Giustizia e Imparzialità

L'intelligenza artificiale (AI) possiede il potenziale per rivoluzionare numerosi settori e migliorare la qualità della vita in molteplici modi. Tuttavia, una delle sfide principali nello sviluppo e nell'implementazione dei sistemi di AI è la presenza di pregiudizi. Questi pregiudizi, definiti come errori sistematici nei processi decisionali che portano a risultati ingiusti, possono emergere da diverse fonti, tra cui la raccolta dei dati, la progettazione degli algoritmi e l'interazione umana.

### Definizione e Tipi di Pregiudizio nell’AI

Nel contesto dell’AI, i pregiudizi possono essere suddivisi in tre categorie principali: pregiudizi nei dati, negli algoritmi e negli utenti. I modelli di apprendimento automatico, una componente essenziale dei sistemi di AI, possono apprendere e replicare i pregiudizi presenti nei dati di addestramento, portando a esiti discriminatori. È cruciale identificare e affrontare questi pregiudizi per garantire l'equità e l'imparzialità dei sistemi di AI.

**Fonti di Pregiudizio nell’AI**

1. **Pregiudizi nei Dati:** Questi pregiudizi si verificano quando i dati utilizzati per addestrare i modelli di apprendimento automatico sono non rappresentativi o incompleti. Ciò può avvenire a causa di fonti di dati pregiudiziali, mancanza di informazioni rilevanti o errori nei dati. Un esempio noto è il sistema di valutazione del rischio COMPAS, utilizzato nel sistema giudiziario statunitense, che ha mostrato pregiudizi contro le persone di colore.
2. **Pregiudizi negli Algoritmi:** Gli algoritmi possono contenere pregiudizi intrinseci derivanti dalle assunzioni e dai criteri utilizzati per prendere decisioni. Ad esempio, il sistema di reclutamento di Amazon ha discriminato le candidate donne poiché era stato addestrato su curriculum prevalentemente maschili.
3. **Pregiudizi degli Utenti:** Gli utenti possono introdurre pregiudizi nei sistemi di AI attraverso i dati di addestramento o interazioni influenzate dai loro stessi pregiudizi. Ad esempio, i sistemi di riconoscimento facciale hanno mostrato tassi di errore più alti per individui di origine africana e asiatica, portando a casi di identificazione errata e arresti ingiusti.

**Strategie di Mitigazione dei Pregiudizi**

Per affrontare e mitigare i pregiudizi nei sistemi di AI, sono stati sviluppati vari approcci:

1. **Augmentazione dei Dataset:** Aggiungere dati più diversificati ai dataset di addestramento per migliorare la rappresentatività e ridurre i pregiudizi. Questo metodo mira a creare un set di dati che rifletta meglio la diversità della popolazione.
2. **Algoritmi Consapevoli del Pregiudizio:** Progettare algoritmi che tengano conto dei diversi tipi di pregiudizio e mirino a minimizzare il loro impatto sugli output del sistema. Questi algoritmi sono sviluppati per riconoscere e correggere le disuguaglianze presenti nei dati di addestramento.
3. **Meccanismi di Feedback degli Utenti:** Raccogliere feedback dagli utenti per identificare e correggere i pregiudizi nel sistema. Questo approccio prevede il coinvolgimento continuo degli utenti nel processo di miglioramento del sistema.

### Equità e Imparzialità nei Modelli Linguistici di Grandi Dimensioni

Il nostro focus si concentra sul principio di equità e imparzialità riguardo ai bias prodotti dai modelli linguistici di grandi dimensioni (LLM). Vi è una crescente preoccupazione che i contenuti generati dagli algoritmi di machine learning (ML) possano discriminare ingiustamente basandosi su attributi demografici personali. Questi modelli, utilizzati nel contesto dell'AI generativa, producono contenuti come testo, immagini o dati e, di conseguenza, sono soggetti a pregiudizi generativi.

Il pregiudizio generativo emerge quando gli output del modello riflettono sproporzionatamente attributi, prospettive o modelli specifici presenti nei dati di addestramento, portando a rappresentazioni distorte o sbilanciate nei contenuti generati. Un modello di generazione di testi addestrato prevalentemente sulla letteratura di autori occidentali, può sovra-rappresentare le norme culturali e gli idiomi occidentali, sottorappresentando o rappresentando erroneamente altre culture. Allo stesso modo, un modello di generazione di immagini addestrato su dataset con scarsa diversità nei ritratti umani può faticare a rappresentare accuratamente una vasta gamma di etnie.

Gli studiosi hanno risposto introducendo numerose definizioni matematiche di equità per testare gli algoritmi, molte delle quali sono in conflitto tra loro. Tuttavia, queste rappresentazioni riduzioniste dell'equità spesso non riflettono le considerazioni di equità nella vita reale, che sono altamente contestuali. Inoltre, le metriche di equità tendono a essere implementate all'interno di toolkit specifici per le valutazioni degli algoritmi, risultando difficili da integrare in una valutazione etica più ampia di un algoritmo. A causa della difficoltà di quantificare i bias rilevanti e di districarli dal risultato di interesse, correggere un bias comporta il rischio di aumentare le imprecisioni delle previsioni.

Gli studiosi che tentano di formalizzare questi criteri in una definizione matematica di equità hanno dovuto affrontare la questione di quale tipo di uguaglianza sia ritenuta giusta. Alcuni assumono che qualsiasi disparità in una data metrica di risultato sia inaccettabile. Altri assumono un campo di gioco livellato: ad esempio, non c'è discriminazione di genere o etnica nel mondo reale che possa influenzare i dati. Se una disparità nelle metriche di equità tra gruppi legalmente protetti sia giusta dipende dal contesto. Etnia e genere possono essere causalmente rilevanti in diagnosi mediche differenziali (ad esempio, anemia falciforme, cancro ovarico) a causa dei diversi meccanismi biologici in questione.

Questa analisi sottolinea l'importanza di considerare il contesto nella valutazione dell'equità e di sviluppare metodologie che possano essere adattate a diverse applicazioni, garantendo che i sistemi di AI siano giusti e imparziali per tutti gli utenti.

### Collegare la Filosofia Etica all'Equità Algoritmica

Le definizioni matematiche di equità, sebbene derivino vagamente da una nozione di egualitarismo, dovrebbero essere calcolate tenendo conto delle sfumature e della specificità contestuale presenti nel discorso filosofico. Rivisitando le metriche di equità, questa sezione collegherà ciascuna metrica alla filosofia etica che l'ha ispirata, affrontando anche i divari tra il lavoro filosofico e ciò che è rappresentato nella formula matematica.

Ad esempio, la massimizzazione dell'accuratezza è soggetta a bias introdotti nel ciclo di sviluppo del modello che possono distorcere le previsioni, il che è particolarmente problematico se i bias riflettono schemi di discriminazione sociale, portando a risultati "non meritati" contrari alla filosofia del merito. La parità demografica è problematica se esistono ragioni legittime dietro l'esito disuguale (ad esempio, reddito disuguale). La metrica dell'equa opportunità, pur sembrando attraentemente simile all'eguaglianza di opportunità di Rawls, non affronta la discriminazione che potrebbe essere già incorporata nei dati. La discriminazione può essere cristallizzata nel set di dati a causa della raccolta di dati distorta (ad esempio, marketing selettivo), etichettatura dei dati distorta (ad esempio, esseri umani che valutano i candidati maschi come più competenti) o decisioni umane distorte che alimentano il sistema (ad esempio, se i tribunali sono più propensi a giudicare colpevoli gli imputati neri). L'eguaglianza di opportunità di Rawls assume anche che le disuguaglianze nel talento innato e nell'ambizione possano portare a risultati disuguali, cosa che non viene affrontata nella parità dei tassi di falsi negativi.

Ogni metrica di equità di gruppo, inclusi gli odds uguali, la parità predittiva positiva e il bilanciamento della classe positiva/negativa, richiede diverse assunzioni sul divario tra lo spazio osservato (caratteristiche) e lo spazio del costrutto (variabili non osservabili): "se c'è un bias strutturale nella pipeline decisionale, nessun meccanismo di equità di gruppo può garantire l'equità". Questo è supportato da una critica alle metriche esistenti di parità di classificazione, in cui gli autori concludono che "nella misura in cui le metriche di errore differiscono tra i gruppi, ciò ci dice di più sulle forme delle distribuzioni del rischio che sulla qualità delle decisioni". In molti domini in cui ci sono preoccupazioni sul bias algoritmico ingiusto, incluso il rischio di credito e l'occupazione, c'è spesso una storia documentata di discriminazione strutturale e sociale, che può influenzare i dati di base attraverso i bias precedentemente discussi.

La sfida dell'equità individuale è: come definire la "somiglianza" che è, ad esempio, indipendente dall’etnia. Quando le caratteristiche predittive sono anche influenzate da caratteristiche protette, la designazione di una misura di "somiglianza" non può essere indipendente da quelle caratteristiche protette. Ad esempio, quale proporzione della disparità di reddito di genere è dovuta alla discriminazione strutturale sul lavoro rispetto alle preferenze lavorative? Alcuni studiosi hanno tentato di incorporare correzioni attive per la disuguaglianza etnica nelle metriche di somiglianza, ma questo dipende fortemente dall'assunzione che la disuguaglianza dovuta alla discriminazione etnica possa essere isolata da altre fonti di disuguaglianza.

Mentre le metriche di equità controfattuale forniscono un'elegante astrazione dell'algoritmo, i meccanismi causali, ad esempio di un'inadempienza su un prestito o sul rischio assicurativo, non sono tipicamente ben compresi. È anche difficile isolare l'impatto di una caratteristica protetta, ad esempio l’etnia, sul risultato, ad esempio il rischio di inadempienza, dalle restanti caratteristiche. L'approccio è anche sensibile alle variabili confondenti non misurate, che possono aggiungere ulteriore bias discriminatorio. I confondenti sono particolarmente difficili da determinare per modelli complessi.

### Esempi concreti di discriminazione dovuta a bias di vario tipo

Per comprendere l’impatto che un modello non correttamente addestrato può causare sul benessere ed autonomia delle persone, è importante raccontare alcuni fatti noti riguardanti modelli non correttamente addestrati e soggetti a discriminazione:

1. **Discriminazione Etnica nei Sistemi di Giustizia Penale:** Uno dei casi più emblematici di bias algoritmico riguarda il sistema COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), utilizzato negli Stati Uniti per valutare il rischio di recidiva dei detenuti. Nel 2016, un'inchiesta di ProPublica ha rivelato che COMPAS tendeva a sovrastimare il rischio di recidiva per gli imputati afroamericani e a sottostimarlo per quelli bianchi, nonostante la precisione complessiva del sistema fosse simile per entrambi i gruppi etnici. Questo bias ha portato a decisioni di custodia preventiva potenzialmente ingiuste, evidenziando il pericolo di affidarsi a sistemi algoritmici senza un'adeguata verifica e comprensione dei loro limiti e pregiudizi intrinseci.
2. **Bias di Genere nei Sistemi di Reclutamento:** Nel 2018, Amazon ha scoperto che il suo sistema di reclutamento basato sull'intelligenza artificiale discriminava sistematicamente le candidate donne. L'algoritmo era stato addestrato sui curriculum vitae inviati ad Amazon nell'ultimo decennio, la maggior parte dei quali proveniva da uomini. Di conseguenza, l'algoritmo penalizzava i curriculum che contenevano parole come "women's" e assegnava punteggi più bassi alle candidature femminili. Questo esempio illustra come i bias storici presenti nei dati di addestramento possano perpetuare e amplificare le disuguaglianze di genere.
3. **Riconoscimento Facciale e Bias Etnico:** I sistemi di riconoscimento facciale utilizzati dalle forze dell'ordine e dalle aziende private hanno mostrato significativi bias etnici. Uno studio del National Institute of Standards and Technology (NIST) ha rilevato che molti algoritmi di riconoscimento facciale hanno tassi di errore più alti per individui di origine africana e asiatica rispetto a quelli di origine europea. Un caso concreto riguarda Robert Julian-Borchak Williams, un uomo afroamericano arrestato erroneamente a Detroit nel 2020 a causa di un errore di identificazione del software di riconoscimento facciale. Questo incidente mette in luce i gravi rischi di affidarsi a tecnologie non sufficientemente accurate e imparziali in contesti critici come l'applicazione della legge.
4. **Assistenti Virtuali e Stereotipi di Genere:** Gli assistenti virtuali come Siri di Apple e Alexa di Amazon hanno mostrato bias di genere nei loro comportamenti e risposte. Per molto tempo, questi assistenti tendevano a rispondere a insulti o commenti sessisti con risposte passive o scherzose, rafforzando stereotipi di genere dannosi. In risposta alle critiche, sia Apple che Amazon hanno aggiornato i loro assistenti virtuali per rispondere in modo più appropriato a tali commenti. Questo esempio evidenzia l'importanza di progettare tecnologie che non perpetuino stereotipi e pregiudizi.
5. **Bias nei Risultati di Ricerca:** Google è stata criticata per i bias nei suoi algoritmi di ricerca, che a volte riflettono pregiudizi esistenti nella società. Ad esempio, una ricerca su "CEO" tendeva a restituire immagini prevalentemente di uomini bianchi, anche se ci sono molti leader aziendali che non rientrano in questa categoria. Questo tipo di bias può rafforzare stereotipi e disuguaglianze preesistenti, influenzando la percezione pubblica di chi è adatto a ricoprire ruoli di leadership.
6. **Traduttori Automatici e Bias di Genere:** I sistemi di traduzione automatica, come Google Translate, hanno mostrato bias di genere. Ad esempio, frasi neutre in lingue che non specificano il genere (come il turco) venivano tradotte in inglese con pronomi di genere che riflettevano stereotipi di genere tradizionali (ad esempio, "He is a doctor" e "She is a nurse" per la stessa frase neutra in turco). Google ha poi implementato modifiche per cercare di mitigare questo problema, offrendo traduzioni che includono entrambi i generi. Questo esempio dimostra come i sistemi di intelligenza artificiale possano riflettere e perpetuare stereotipi culturali presenti nei dati di addestramento.

Tipo di Pregiudizio e Descrizione

|  |  |  |
| --- | --- | --- |
| Tipo di Pregiudizio | Descrizione | Esempi |
| Pregiudizio di campionamento | Si verifica quando i dati di addestramento non sono rappresentativi della popolazione a cui si rivolgono, portando a scarse prestazioni e previsioni pregiudizievoli per alcuni gruppi. | Un algoritmo di riconoscimento facciale addestrato principalmente su individui bianchi che funziona male su persone di altre etnie. |
| Pregiudizio algoritmico | Risulta dalla progettazione e implementazione dell'algoritmo che può dare priorità a certi attributi, portando a esiti ingiusti. | Un algoritmo che dà priorità all'età o al genere, portando a esiti ingiusti nelle decisioni di assunzione. |
| Pregiudizio di rappresentazione | Si verifica quando un set di dati non rappresenta accuratamente la popolazione che intende modellare, portando a previsioni inaccurate. | Un set di dati medici che sottorappresenta le donne, portando a diagnosi meno accurate per le pazienti femminili. |
| Pregiudizio di conferma | Si materializza quando un sistema di IA viene utilizzato per confermare pregiudizi o credenze preesistenti dei suoi creatori o utenti. | Un sistema di IA che predice il successo dei candidati al lavoro basandosi sui pregiudizi del responsabile delle assunzioni. |
| Pregiudizio di misurazione | Emerge quando la raccolta o la misurazione dei dati sovra o sottorappresenta sistematicamente certi gruppi. | Un sondaggio che raccoglie più risposte dai residenti urbani, portando a una sottorappresentazione delle opinioni rurali. |
| Pregiudizio di interazione | Si verifica quando un sistema di IA interagisce con gli esseri umani in modo pregiudizievole, risultando in un trattamento ingiusto. | Un chatbot che risponde diversamente a uomini e donne, risultando in una comunicazione pregiudizievole. |
| Pregiudizio generativo | Si verifica nei modelli di IA generativa, come quelli utilizzati per creare dati sintetici, immagini o testo. Il pregiudizio generativo emerge quando gli output del modello riflettono in modo sproporzionato attributi, prospettive o schemi specifici presenti nei dati di addestramento, portando a rappresentazioni distorte o sbilanciate nei contenuti generati. | Un modello di generazione di testo addestrato prevalentemente sulla letteratura di autori occidentali potrebbe sovrarappresentare le norme e gli idiomi culturali occidentali, sottorappresentando o travisando altre culture. Allo stesso modo, un modello di generazione di immagini addestrato su set di dati con diversità limitata nei ritratti umani potrebbe avere difficoltà a rappresentare accuratamente un'ampia gamma di etnie. |

### Fonti di Bias nell'IA

I sistemi di Intelligenza Artificiale (IA) sono progettati da esseri umani, i quali possono essere soggetti a bias. Quando riconosciamo i nostri bias e i loro impatti sulle decisioni, possiamo creare risultati migliori nelle nostre vite. Con l'emergere e l'adozione diffusa dei sistemi di IA in vari settori industriali, la necessità di comprendere, rilevare e mitigare il bias nelle applicazioni IA ha attirato un notevole interesse. Il bias indesiderato nell'IA presenta numerosi potenziali rischi negativi, sia per l'istituzione implementante che per la sua clientela. Risultati errati dei dati, perdita di fiducia da parte dei dipendenti e dei clienti e impatti negativi sulla reputazione e sui profitti aziendali sono solo alcuni degli effetti indesiderati sulle organizzazioni che non affrontano il bias negativo nei loro sistemi di IA. È stato osservato che il bias può causare una prestazione subottimale dell'IA nella sua applicazione, come la discriminazione dei potenziali clienti. Questo tipo di comportamento del sistema potrebbe portare a un aumento delle cause legali, a una maggiore sorveglianza e a un'erosione del valore di mercato di un'azienda. Per costruire una difesa efficace contro questi rischi, è essenziale prima comprendere le fonti del bias nell'IA.

Nella maggior parte delle applicazioni di IA, il bias può derivare da molte fonti. Tra queste fonti comuni di bias nelle applicazioni di IA troviamo:

1. **Bias dal Disegno della Ricerca**: I bias derivanti dal disegno della ricerca, noti anche come bias umani, riflettono i bias degli sviluppatori del sistema di IA. L'obiettivo e l'implementazione dell'IA possono influenzare negativamente i gruppi sottorappresentati. Questo tipo di bias può essere introdotto durante la raccolta e il filtraggio dei dati, la selezione soggettiva delle caratteristiche o durante la valutazione del modello utilizzando misure e tecniche di valutazione specificamente progettate. Studi recenti hanno scoperto che un'alta percentuale degli studi che utilizzano modelli predittivi di IA presentava un alto rischio di bias, con fattori contributivi principali come la cattiva gestione dei dati mancanti e il mancato affrontare l'overfitting.
2. **Bias dai Dati di Addestramento**: I bias dai dati di addestramento possono derivare da due condizioni. In primo luogo, i dati possono essere manipolati per distorcere i risultati, come nel caso dell'addestramento su dati pubblici o facilmente modificabili (es. post sui social media, articoli wiki) non accuratamente curati. In secondo luogo, se il set di dati utilizzato per addestrare un modello non è rappresentativo della popolazione target, si introdurranno errori o previsioni distorte quando applicato a nuovi dati di quella popolazione.
3. **Bias dalle Rappresentazioni degli Input**: Conosciuti come bias di rappresentazione, questi derivano da come i dati vengono rappresentati. Ad esempio, le rappresentazioni degli input, come gli embedding di parole o frasi, tendono a catturare gli atteggiamenti sociali e a mostrare bias semantici. Un esempio è l'associazione distorta tra generi e professioni, come infermieri femminili e medici maschili.
4. **Bias dall'Architettura del Modello di IA**: L'architettura del modello può portare a bias algoritmici, che consistono nel perpetuare le disuguaglianze esistenti tra gruppi svantaggiati. Questa forma di bias è spesso nascosta agli osservatori esterni, rendendola difficile da affrontare. La mancanza di una misura standard del bias complica ulteriormente la situazione, con misure comuni come pari opportunità equalizzate, parità statistica e parità predittiva che si applicano solo in circostanze specifiche.
5. **Bias dall'Uso nel Mondo Reale**: I bias possono sorgere dopo la distribuzione del sistema di IA e il suo utilizzo in contesti reali. Questi bias possono emergere quando un sistema funziona senza bias in un contesto ma produce risultati distorti in un altro, o quando un modello adattivo impara bias dai suoi utenti nel tempo.

A causa della loro recente popolarità, i grandi modelli linguistici (LLM) come ChatGPT e GPT-4 sono stati un importante centro di attenzione riguardo al bias. Un modello linguistico è un modello di apprendimento automatico di una lingua naturale utilizzato per compiti come la risposta a domande, la sintesi e la classificazione del testo. Gli LLM, come qualsiasi prodotto di IA, sono suscettibili ai problemi di bias in ogni fase dello sviluppo, in particolare nell'acquisizione e nell'utilizzo dei dati di addestramento. Anche se gli sviluppatori di ChatGPT e altri LLM prendono misure per rimuovere dati di bassa qualità, espliciti o potenzialmente dannosi, questi modelli possono comunque esprimere bias appresi dai loro dati di addestramento, rendendo la misurazione e la mitigazione di questo bias una sfida significativa per le ricerche future. Un'ulteriore considerazione riguarda l'autenticità dei dati di addestramento utilizzati, poiché i set di dati pubblici potrebbero essere distorti o contenere “fake news”.

# Bias nelle Intelligenze Artificiali

|  |  |  |
| --- | --- | --- |
| Nome del Bias | Descrizione | Esempio |
| Bias di conferma | L'AI tende a favorire informazioni che confermano le proprie ipotesi o pregiudizi preesistenti. | Un sistema di raccomandazione che suggerisce continuamente articoli simili a quelli già letti dall'utente. |
| Bias di selezione | Errori derivanti dalla selezione di un campione non rappresentativo per l'addestramento. | Un modello di riconoscimento facciale addestrato solo su immagini di persone di un certo gruppo etnico. |
| Bias di stereotipo | L'AI perpetua stereotipi esistenti basati su dati di addestramento parziali o discriminatori. | Un assistente virtuale che associa automaticamente determinate professioni a un genere specifico. |
| Bias di segregazione | Separazione in gruppi che risultano discriminati nei processi decisionali dell'AI. | Un algoritmo di prestiti che tende a negare richieste da quartieri con predominanza di minoranze etniche. |
| Bias di omogeneità | L'AI generalizza eccessivamente dati da gruppi non sufficientemente rappresentati. | Un sistema di diagnosi medica che funziona meglio per un genere rispetto all'altro. |
| Bias di disponibilità | Influenza di dati facilmente disponibili o recenti sul modello decisionale dell'AI. | Un motore di ricerca che favorisce risultati più recenti ma meno pertinenti rispetto a query passate. |
| Bias di ancoraggio | L'AI si basa eccessivamente sui primi dati osservati o su un punto di riferimento iniziale. | Un sistema di pricing che si basa su valori iniziali senza adattarsi a nuovi dati di mercato. |
| Bias di interazione | Bias introdotti attraverso l'interazione con utenti umani che influenzano l'addestramento. | Un chatbot che diventa offensivo o discriminatorio in risposta alle interazioni degli utenti. |
| Bias di conferma sociale | Influenza dei comportamenti collettivi o di gruppo sul modello decisionale dell'AI. | Un algoritmo di raccomandazione che promuove contenuti già popolari, ignorando le preferenze individuali. |
| Bias di rappresentazione | Distorsione derivante dalla rappresentazione errata o parziale dei dati di addestramento. | Un sistema di traduzione che ha difficoltà con linguaggi o dialetti meno rappresentati nei dati di addestramento. |
| Bias di sopravvivenza | Distorsione causata dal focalizzarsi solo su dati o casi che hanno superato un certo processo. | Un'analisi di successo aziendale che ignora le aziende fallite, risultando in un quadro distorto del successo. |
| Bias di percezione selettiva | Focalizzarsi su determinate caratteristiche o dati ignorando altri aspetti rilevanti. | Un modello di valutazione del rischio che considera solo determinati fattori, tralasciandone altri importanti. |
| Bias di osservazione | Distorsione derivante dal modo in cui vengono raccolti i dati di osservazione. | Dati medici raccolti solo in ospedali di una certa area geografica che non rappresentano altre popolazioni. |
| Bias di reporting | Distorsione derivante dal fatto che alcuni eventi o risultati sono più probabilmente riportati. | Dati sulla criminalità basati su denunce di polizia, che possono non riflettere la realtà completa dei crimini. |
| Bias di automantenimento | Sistemi che perpetuano i propri pregiudizi attraverso feedback loop. | Un algoritmo di raccomandazione che continua a suggerire solo certi tipi di contenuti basati su feedback iniziali. |
| Bias di design | Influenza delle scelte di design sul comportamento e i risultati del sistema AI. | Un'interfaccia utente che influenza gli utenti a fare scelte che il designer preferisce. |
| Bias di interpretazione | Distorsione nell'interpretazione dei risultati basata su pregiudizi preesistenti. | Un'analisi dei dati che interpreta i risultati in modo da confermare le ipotesi iniziali. |
| Bias di conferma selettiva | Tendenza a cercare, interpretare e ricordare informazioni che confermano le proprie convinzioni. | Un'AI che seleziona solo dati che supportano le sue decisioni precedenti. |
| Bias di frammentazione | Errori derivanti dalla frammentazione dei dati in piccoli segmenti non rappresentativi. | Un modello che prende decisioni basate su piccoli campioni di dati, perdendo il contesto generale. |
| Bias di attributi | Distorsione derivante dall'assegnazione impropria di attributi a determinati gruppi o individui. | Un sistema di valutazione delle performance che favorisce attributi non correlati con le reali competenze. |
| Bias di comparazione | Distorsione derivante dal confronto inappropriato tra diversi gruppi o individui. | Un modello che valuta le prestazioni degli studenti confrontandoli solo con studenti di una scuola specifica. |
| Bias di inclusione/esclusione | Errori derivanti dall'inclusione o esclusione di certi dati o categorie. | Un modello che esclude dati di minoranze etniche, portando a decisioni non rappresentative. |
| Bias di sistematicità | Distorsione derivante da errori sistematici presenti nei dati di addestramento. | Un modello predittivo che riflette errori di misurazione sistematica nei dati raccolti. |

### Misurare il Bias dell'IA

Comprendere i problemi di bias che possono sorgere nelle applicazioni di Intelligenza Artificiale (IA) è solo il primo passo; la sfida successiva riguarda la misurazione di questi bias. In alcuni casi, il bias di un sistema di IA può essere quantificato utilizzando misure quantitative generali come pari opportunità equalizzate, parità statistica, parità predittiva o equità controfattuale. Tuttavia, queste misure potrebbero non essere applicabili in tutte le situazioni, inclusi quei casi in cui sono necessarie valutazioni uniche del bias per quantificare il bias di interesse.

Molte misure di bias sono state sviluppate per quantificare il bias nelle rappresentazioni degli input, come gli embedding di parole, per determinare l'efficacia delle tecniche di mitigazione. Ad esempio, il Word-Embedding Association Test (WEAT) e i suoi derivati, come il SEAT, sono stati ampiamente utilizzati in diversi studi che investigano metodi di mitigazione del bias. WEAT è stato creato per valutare il bias nelle rappresentazioni semantiche delle parole nell'IA, che rappresentano le parole come vettori basati sui contesti testuali in cui si trovano le parole. Questa misura funziona considerando due insiemi di parole target e due insiemi di parole attributo, con l'ipotesi nulla che non esistano differenze tra gli insiemi di parole target e le loro somiglianze relative agli insiemi di parole attributo. Il bias viene quindi quantificato calcolando la probabilità che una permutazione di parole attributo produrrebbe la differenza osservata nelle medie campionarie, determinando così l'improbabilità dell'ipotesi nulla. WEAT può essere utilizzato per determinare rapidamente le differenze di bias tra gruppi demografici.

Diverse misure sono state create per valutare forme specifiche di bias, inclusi bias stereotipici ed etnici. Ad esempio, è stato sviluppato un set di dati chiamato StereoSet, che è un grande set di dati di linguaggio naturale creato per valutare i bias stereotipici in genere, professione, razza e religione. Questo set di dati è accoppiato con una misura del bias stereotipico, il Test di Associazione di Contesto (CAT), che quantifica la capacità di modellazione linguistica e il bias stereotipico dei modelli linguistici pre-addestrati. In CAT, al modello viene fornita una frase di contesto contenente un gruppo target, e deve associare questo contesto a una risposta stereotipica, anti-stereotipica o non correlata. Con questo, le associazioni stereotipiche e anti-stereotipiche quantificano il bias stereotipico, mentre le associazioni non correlate quantificano la capacità di modellazione linguistica.

Un'altra misura del bias esamina il bias etnico dipendente dalla lingua. Per osservare e quantificare questo bias, è stato sviluppato il punteggio di Bias Categoriale (CB), che determina il grado di varianza nella probabilità che il modello linguistico restituisca un nome di paese dato un attributo in una frase senza alcun indizio rilevante. Utilizzando il punteggio CB, è stato valutato il bias etnico dipendente dalla lingua in modelli monolingue e multilingue, scoprendo che il modello inglese aveva il livello più basso di bias etnico e che il modello multilingue produceva punteggi CB più bassi per diverse lingue testate.

Altre misure di bias includono il Test di Coerenza degli Embedding (ECT) e varie misure di associazione e inferenza del linguaggio naturale. Ad esempio, il punteggio ECT determina se gruppi di parole hanno associazioni stereotipiche calcolando il coefficiente di Spearman di elenchi di embedding di attributi ordinati in base alle loro somiglianze con l'embedding target. Misure di associazione come l'informazione reciproca normalizzata (nPMI) possono quantificare i bias appresi da un modello in assenza di verità di base, utilizzando le previsioni di un modello di classificazione per un'immagine come un insieme di etichette. Inoltre, i punteggi di inferenza del linguaggio naturale possono quantificare l'effetto che le associazioni distorte hanno sulle decisioni prese a valle, utilizzando coppie di frasi costruite in modo neutrale che differiscono solo nel soggetto. Queste misure, insieme a WEAT, CAT e CB, possono essere utilizzate per quantificare varie forme di bias nei sistemi di IA, dimostrando che la misurazione del bias dell'IA continua a presentare una sfida significativa.

Negli ultimi mesi, l'interesse per la misurazione e la valutazione del bias degli LLM, come ChatGPT, è cresciuto. Alcuni approcci per identificare il bias in tali sistemi includono la conduzione di audit regolari sui risultati del modello, nonché l'applicazione di misure generali di bias come pari opportunità equalizzate e parità statistica. Tuttavia, l'uso di tali misure potrebbe non essere applicabile o desiderabile in ogni situazione; pertanto, è necessaria una ricerca di metodi alternativi per identificare e misurare il bias. È stato esplorato il problema della misurazione del bias in ChatGPT valutando questo modello e altri modelli con benchmark specifici per il bias, trovando che ChatGPT potrebbe esprimere meno bias rispetto ad altri LLM. Tuttavia, questi modelli continuano a mostrare bias nei loro output, evidenziando la necessità di mitigare il bias in questi e altri sistemi di IA.

### Misurare il Bias dell'IA

Oltre alla misurazione del bias per un sistema di IA, è fondamentale considerare set di dati di benchmark che possono essere utilizzati per quantificare il bias o valutare l'impatto di una tecnica di mitigazione del bias sulle prestazioni del modello.

#### Set di Dati di Misurazione del Bias

A volte, set di dati specializzati vengono sviluppati insieme a nuove misure del bias. Uno di questi set di dati è StereoSet, creato per quantificare il bias stereotipico nei modelli linguistici pre-addestrati. Questo set di dati di linguaggio naturale può essere utilizzato per valutare il bias stereotipico in quattro domini: genere, professione, razza e religione. Per ciascuno di questi domini, sono stati selezionati termini target rappresentanti un gruppo sociale. Ad esempio, termini come “medico” e “allenatore” per il dominio delle professioni, e termini come “patrigno” e “nonna” per il dominio del genere. Sono stati sviluppati gruppi di frasi con un focus su questi termini target per compiti intra-frase e inter-frase, includendo frasi stereotipate, anti-stereotipate e non correlate.

Altri set di dati specializzati per misurare il bias nell'IA includono WinoBias, BBQ e BOLD. WinoBias è progettato per identificare meglio il bias di genere nei sistemi di risoluzione delle coreferenze, modelli di IA utilizzati per trovare tutte le espressioni che si riferiscono alla stessa entità in un testo. Questo set di dati contiene frasi etichettate come “pro-stereotipiche” o “anti-stereotipiche” in base al fatto che il pronome corrisponda al genere dominante nella professione.

BBQ è sviluppato per scoprire e quantificare i bias sociali nei risultati dei modelli per il compito di risposta alle domande. Questo set di dati valuta diversi bias, inclusi quelli relativi all'età, al genere, alla razza e allo stato socioeconomico. BBQ determina quanto fortemente le risposte del modello riflettono i bias sociali e se questi bias prevalgono su una scelta di risposta corretta.

BOLD è progettato per quantificare i bias sociali nelle generazioni di testo aperte prodotte da modelli di IA generativa, come i chatbot. Questo set di dati valuta il bias in cinque domini: professione, genere, razza, religione e ideologia politica. Con misure di tossicità, norme psicolinguistiche e polarità di genere del testo, BOLD può essere utilizzato per valutare e quantificare i bias sociali presenti nelle generazioni di testo aperte da più angolazioni.

#### Set di Dati di Valutazione dell'Impatto della Mitigazione del Bias

Oltre ai set di dati specializzati che quantificano il bias nei sistemi di IA, diversi set di dati di benchmark sono spesso utilizzati per determinare l'impatto delle tecniche di mitigazione del bias sulle prestazioni di un modello. Tra questi, i più comuni includono il Corpus of Linguistic Acceptability (CoLA) e lo Stanford Sentiment Treebank (SST-2). Ciascuno di questi set di dati fa parte del General Language Understanding Evaluation (GLUE) benchmark, una raccolta di strumenti per valutare le prestazioni dei modelli su un insieme diversificato di compiti di comprensione del linguaggio naturale.

CoLA e SST-2 possono essere utilizzati per valutare le prestazioni dei sistemi di IA per compiti a frase singola. Il set di dati CoLA si concentra sull'accettabilità grammaticale e consiste in giudizi derivati dalla teoria linguistica. Contiene sequenze di parole etichettate come frasi grammaticali o non grammaticali e utilizza la misura del coefficiente di correlazione di Matthews per la valutazione. Il set di dati SST-2 si concentra sul compito di prevedere il sentimento di una data frase, composto da frasi etichettate con un sentimento positivo o negativo.

Confrontando le prestazioni del modello prima e dopo l'implementazione di una tecnica di mitigazione del bias, l'impatto relativo di questa tecnica può essere rappresentato dalle differenze nelle prestazioni sui set di dati CoLA e SST-2. Ad esempio, questi set di dati sono stati utilizzati per valutare le prestazioni di due tecniche di debiasing: Sent-Debias e Auto-Debias. Poiché queste tecniche mitigano il bias modificando il metodo di embedding delle frasi del modello e i parametri di messa a punto, rispettivamente, c'è preoccupazione che possano influenzare le prestazioni del modello o la capacità di comprensione del linguaggio. Gli studi hanno dimostrato che queste tecniche causano diminuzioni minime o nulle nelle capacità di comprensione del linguaggio dei modelli, dimostrando l'importanza di questi set di dati per il benchmarking del bias nell'IA.

#### Linee Guida per Ridurre il Bias nei Sistemi di IA

Il bias può derivare sia da ciò che conosciamo, come preferenze e associazioni, sia da ciò che non conosciamo, come dati incompleti o mancanti. Alcune linee guida comuni per ridurre il bias nei sistemi di IA includono:

1. **Creare un Obiettivo Ben Definito**: Lavorare a stretto contatto con le parti interessate per definire chiaramente l'obiettivo del sistema. Bias intrinseci del team di sviluppo o dati di addestramento limitati possono causare false ipotesi che potrebbero influenzare negativamente i gruppi sottorappresentati. Ad esempio, gli sviluppatori potrebbero assumere che solo un certo gruppo demografico sia interessato a un prodotto, ignorando altri gruppi che potrebbero trarne beneficio.
2. **Rivedere i Dati di Addestramento e di Input**: Analizzare i dati per segni di bias. Dati inaccurati, mancanti o manipolati possono causare risultati distorti. Ad esempio, se l'IA viene addestrata su dati pubblici come post sui social media o articoli wiki, c'è il rischio che questi dati non siano autentici o validati, distorcendo i risultati e influenzando negativamente certi gruppi. È importante considerare come i dati potrebbero essere manipolati e abilitare salvaguardie per proteggersi da tale manipolazione.
3. **Utilizzare Modelli Spiegabili e Interpretabili**: Fornire trasparenza nello sviluppo e nei metodi dell'algoritmo può ridurre il bias algoritmico. Divulgare gli input, i parametri e gli output dell'algoritmo aiuta a comprendere come l'algoritmo prende le decisioni. Utilizzare modelli interpretabili e ispezionare la logica decisionale attraverso la spiegabilità del modello può ulteriormente ridurre il bias.

#### Tecnologie di Debiasing

Diverse tecnologie di debiasing sono state introdotte per ridurre il bias negli algoritmi e nei sistemi di IA esistenti. Queste tecnologie si concentrano sulla debiasing delle distribuzioni dei dati, delle rappresentazioni degli embedding e dei parametri di messa a punto.

1. **Riaddestramento con Set di Dati Bilanciati**: Una strategia comune è riaddestrare il modello con un set di dati più bilanciato. Un metodo di debiasing senza perdita di informazioni, sviluppato da Zhou, Kantarcioglu e Clifton, si concentra sulla scarsità di dati nei gruppi svantaggiati. Questo metodo sovracampiona i gruppi sottorappresentati generando dati sintetici per aumentare la loro rappresentazione e eliminare il bias intrinseco nei dati, mantenendo le prestazioni del modello.
2. **Sent-Debias**: Proposto da Liang et al., Sent-Debias è un metodo di debiasing delle rappresentazioni delle frasi in quattro passaggi: definire le parole che mostrano attributi di bias, contestualizzare queste parole in frasi con attributi di bias, stimare il sottospazio del bias delle rappresentazioni delle frasi e debiasing delle frasi generali rimuovendo la proiezione su questo sottospazio del bias. Sent-Debias ha dimostrato di produrre rappresentazioni delle frasi con bias limitato.
3. **Auto-Debias**: Proposto da Guo, Yang e Abbasi, Auto-Debias è un metodo per mitigare i bias sociali nei modelli linguistici pre-addestrati in due fasi: creare automaticamente prompt distorti massimizzando il disaccordo tra i completamenti del modello linguistico mascherato, e utilizzare questi prompt per mettere a punto il modello minimizzando il disaccordo tra i suoi completamenti. Auto-Debias ha ottenuto una prestazione di debiasing elevata rispetto ad altri metodi.

#### Sfide nei Grandi Modelli Linguistici (LLM)

A causa delle loro dimensioni e complessità, i grandi modelli linguistici (LLM) come ChatGPT e GPT-4 presentano sfide uniche riguardo alla mitigazione del bias. Mitigare il bias in questi modelli richiede uno sforzo collaborativo tra sviluppatori di IA, utenti e comunità interessate. Alcune potenziali strategie includono:

1. **Coinvolgere le Comunità Svantaggiate**: Durante lo sviluppo del modello, coinvolgere le comunità svantaggiate per garantire che le loro esigenze siano considerate.
2. **Collaborare con Esperti Multidisciplinari**: Lavorare con esperti di varie discipline per identificare e mitigare il bias.
3. **Considerare il Feedback degli Utenti**: Valutare regolarmente i risultati del modello attraverso il feedback degli utenti.
4. **Trasparenza e Apertura**: Essere trasparenti riguardo alle metodologie, alle fonti dei dati e ai potenziali bias del modello.
5. **Stabilire Partnership**: Creare partnership tra ricercatori e parti esterne per condividere conoscenze e migliori pratiche.

Con queste strategie, insieme alle tecniche di debiasing discusse, possono essere sviluppati sistemi di IA più equi e inclusivi, riducendo il bias verso individui o gruppi svantaggiati.

### **Implicazioni Etiche dell'Intelligenza Artificiale Pregiudiziale**

L'uso di intelligenza artificiale (AI) pregiudiziale solleva numerose implicazioni etiche che devono essere attentamente considerate. Una delle principali preoccupazioni riguarda il potenziale di discriminazione nei confronti di individui o gruppi basati su fattori quali etnia, genere, età o disabilità . Quando i sistemi di AI sono affetti da pregiudizi, possono perpetuare le disuguaglianze esistenti e rafforzare la discriminazione contro i gruppi emarginati. Questo è particolarmente allarmante in aree sensibili come la sanità, dove i sistemi di AI pregiudiziali possono causare disuguaglianze nell'accesso alle cure o arrecare danni ai pazienti .

Un'altra questione etica cruciale è la responsabilità degli sviluppatori, delle aziende e dei governi nel garantire che i sistemi di AI siano progettati e utilizzati in modo equo e trasparente. Se un sistema di AI è pregiudiziale e produce esiti discriminatori, la responsabilità ricade non solo sul sistema stesso, ma anche su coloro che lo hanno creato e implementato . Pertanto, è essenziale stabilire linee guida etiche e quadri normativi che rendano responsabili tutti coloro che sviluppano e utilizzano i sistemi di AI per qualsiasi risultato discriminatorio.

Inoltre, l'uso di sistemi di AI pregiudiziali può minare la fiducia del pubblico nella tecnologia, portando a una diminuzione dell'adozione e persino al rifiuto delle nuove tecnologie. Questo può avere serie implicazioni economiche e sociali, poiché i potenziali benefici dell’AI potrebbero non essere realizzati se le persone non si fidano della tecnologia o la percepiscono come uno strumento di discriminazione.

È anche importante considerare l'impatto dell’AI pregiudiziale sull'autonomia e l'agenzia umana. Quando i sistemi di AI sono pregiudiziali, possono limitare le libertà individuali e rafforzare le dinamiche di potere sociali. Ad esempio, un sistema di AI utilizzato nei processi di assunzione potrebbe escludere in modo sproporzionato i candidati provenienti da gruppi emarginati, limitando la loro capacità di accedere a opportunità lavorative e di contribuire alla società.

Affrontare le implicazioni etiche dell’AI pregiudiziale richiederà uno sforzo concertato da parte di tutti gli stakeholder coinvolti, inclusi sviluppatori, politici e la società in generale. Sarà necessario sviluppare linee guida etiche e quadri normativi che promuovano l'equità, la trasparenza e la responsabilità nello sviluppo e nell'uso dei sistemi di AI . Inoltre, sarà fondamentale avviare discussioni critiche sull'impatto dell’AI sulla società e permettere agli individui di partecipare alla modellazione del futuro dell’AI in modo responsabile ed etico.

### Panoramica delle Strategie di Mitigazione del Pregiudizio relative ad un certo Bias

Mitigare il pregiudizio nell'intelligenza artificiale rappresenta una sfida complessa e diversificata, ma sono stati proposti diversi approcci per affrontare questo problema. Un metodo comune consiste nel pre-processare i dati utilizzati per addestrare i modelli di AI, al fine di garantirne la rappresentatività dell'intera popolazione, inclusi i gruppi storicamente emarginati. Ciò può comportare tecniche come l'oversampling, l'undersampling o la generazione di dati sintetici. Ad esempio, uno studio di Buolamwini e Gebru ha dimostrato che l'oversampling di individui con pelle più scura ha migliorato l'accuratezza degli algoritmi di riconoscimento facciale per questo gruppo.

Il pre-processamento dei dati richiede l'identificazione e l'eliminazione dei pregiudizi presenti nei dati prima dell'addestramento del modello. Questo può essere realizzato mediante tecniche come l'augmentazione dei dati, che comporta la creazione di punti dati sintetici per aumentare la rappresentazione dei gruppi sottorappresentati, o attraverso la de-biasing avversariale, che comporta l'addestramento del modello per essere resistente a specifici tipi di pregiudizio. Documentare tali pregiudizi nei dataset e le procedure di augmentazione è di fondamentale importanza.

Un altro approccio per mitigare il pregiudizio nell'AI consiste nella selezione attenta dei modelli utilizzati per analizzare i dati. I ricercatori hanno proposto di utilizzare metodi di selezione dei modelli che prioritizzino l'equità, come quelli basati sull'equità di gruppo o sull'equità individuale. Ad esempio, uno studio di Kamiran e Calders ha proposto un metodo per selezionare i classificatori che raggiungono la parità demografica, garantendo che gli esiti positivi e negativi siano distribuiti equamente tra diversi gruppi demografici.

Inoltre, tecniche di selezione dei modelli che prioritizzano l'equità e mitigano il pregiudizio possono includere la regolarizzazione, che penalizza i modelli per fare previsioni discriminatorie, o i metodi di ensemble, che combinano più modelli per ridurre il pregiudizio.

Le decisioni di post-processamento rappresentano un ulteriore approccio per mitigare il pregiudizio nell'IA. Questo comporta l'aggiustamento dell'output dei modelli di IA per rimuovere il pregiudizio e garantire l'equità. Ad esempio, i ricercatori hanno proposto metodi di post-processamento che aggiustano le decisioni prese da un modello per raggiungere pari opportunità, garantendo che i falsi positivi e i falsi negativi siano distribuiti equamente tra diversi gruppi demografici.

In sintesi, la mitigazione del pregiudizio nell'IA richiede un approccio integrato che coinvolga il pre-processamento dei dati, la selezione dei modelli e le decisioni di post-processamento, supportato da una documentazione accurata delle procedure e dei pregiudizi identificati.

Mentre questi approcci sono promettenti per mitigare il pregiudizio nell'intelligenza artificiale, presentano anche limitazioni e sfide. Ad esempio, il pre-processamento dei dati può essere un processo lungo e potrebbe non essere sempre efficace, specialmente se i dati utilizzati per addestrare i modelli sono già pregiudiziali. Inoltre, i metodi di selezione dei modelli possono essere limitati dalla mancanza di consenso su cosa costituisca l'equità, e i metodi di post-processamento possono essere complessi e richiedere grandi quantità di dati aggiuntivi. Pertanto, è cruciale continuare a esplorare e sviluppare nuovi approcci per mitigare il pregiudizio nell'IA.

Nel campo dell'intelligenza artificiale generativa, affrontare il pregiudizio è ancora più impegnativo poiché richiede una strategia olistica. Questo inizia con il pre-processamento dei dati per garantire diversità e rappresentatività. Ciò comporta la raccolta e l'inclusione deliberata di fonti di dati diverse che riflettono la varietà dell'esperienza umana, prevenendo così la sovra-rappresentazione di qualsiasi singolo gruppo demografico nei dataset di addestramento. La selezione dei modelli deve poi prioritizzare gli algoritmi che sono trasparenti e capaci di rilevare quando stanno generando output pregiudiziali. Tecniche come l'addestramento avversariale, dove i modelli sono continuamente testati contro scenari progettati per rivelare pregiudizi, possono essere utili. Il post-processamento comporta la valutazione critica del contenuto generato dall'IA e, se necessario, l'aggiustamento degli output per correggere i pregiudizi. Questo potrebbe includere l'uso di filtri aggiuntivi o tecniche di apprendimento trasferibile per affinare ulteriormente i modelli. Audit regolari, monitoraggio continuo e l'incorporazione di loop di feedback sono essenziali per garantire che i sistemi di IA generativa rimangano equi ed equitativi nel tempo. Questi sforzi devono essere sostenuti da un impegno verso i principi etici dell'IA, coinvolgendo attivamente team diversificati nello sviluppo dell'IA e promuovendo la collaborazione interdisciplinare per affrontare e mitigare efficacemente i pregiudizi nell'IA.

Inoltre, l'implementazione di questi approcci richiede una considerazione attenta delle implicazioni etiche e sociali. Ad esempio, l'aggiustamento delle previsioni del modello per garantire l'equità può comportare compromessi tra diverse forme di pregiudizio e può avere conseguenze non intenzionali sulla distribuzione degli esiti per diversi gruppi.

Approcci per Mitigare i Pregiudizi e Considerazioni Etiche

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Approccio | Descrizione | Esempi | Limitazioni e Sfide | Considerazioni Etiche |
| Pre-elaborazione dei Dati | Coinvolge l'identificazione e la risoluzione dei pregiudizi nei dati prima dell'addestramento del modello. Tecniche come il sovracampionamento, il sottocampionamento o la generazione di dati sintetici vengono utilizzate per garantire che i dati siano rappresentativi dell'intera popolazione, inclusi i gruppi storicamente marginalizzati. | 1. Sovracampionamento degli individui con pelle più scura in un dataset di riconoscimento facciale. 2. Aumento dei dati per incrementare la rappresentanza nei gruppi sottorappresentati. 3. Adversarial debiasing per addestrare il modello a essere resiliente a tipi specifici di pregiudizio. | 1. Processo che richiede tempo. 2. Potrebbe non essere sempre efficace, soprattutto se i dati utilizzati per addestrare i modelli sono già pregiudizievoli. | 1. Potenziale sovra- o sottorappresentazione di certi gruppi nei dati, che può perpetuare pregiudizi esistenti o crearne di nuovi. 2. Preoccupazioni sulla privacy legate alla raccolta e all'utilizzo dei dati, in particolare per i gruppi storicamente marginalizzati. |
| Selezione del Modello | Si concentra sull'utilizzo di metodi di selezione dei modelli che danno priorità all'equità. I ricercatori hanno proposto metodi basati sull'equità di gruppo o sull'equità individuale. Le tecniche includono la regolarizzazione, che penalizza i modelli per fare previsioni discriminatorie, e i metodi di ensemble, che combinano più modelli per ridurre il pregiudizio. | 1. Selezione di classificatori che raggiungono la parità demografica. 2. Utilizzo di metodi di selezione dei modelli basati sull'equità di gruppo o sull'equità individuale. 3. Regolarizzazione per penalizzare le previsioni discriminatorie. 4. Metodi di ensemble per combinare più modelli e ridurre il pregiudizio. | Limitato dalla possibile mancanza di consenso su cosa costituisca l'equità. | 1. Bilanciare l'equità con altre metriche di performance, come l'accuratezza o l'efficienza. 2. Potenziale per i modelli di rinforzare stereotipi o pregiudizi esistenti se i criteri di equità non sono attentamente considerati. |
| Post-elaborazione delle Decisioni | Coinvolge la regolazione dell'output dei modelli di IA per rimuovere il pregiudizio e garantire l'equità. I ricercatori hanno proposto metodi che regolano le decisioni prese da un modello per raggiungere odds equalizzati, garantendo che i falsi positivi e i falsi negativi siano distribuiti equamente tra i diversi gruppi demografici. | Metodi di post-elaborazione che raggiungono odds equalizzati. | Può essere complesso e richiedere grandi quantità di dati aggiuntivi. | 1. Compromessi tra diverse forme di pregiudizio quando si regolano le previsioni per l'equità. 2. Conseguenze non intenzionali sulla distribuzione degli esiti per diversi gruppi. |

### Equità nell'IA

L'equità nell'Intelligenza Artificiale (IA) è un tema critico che ha ricevuto notevole attenzione sia nei circoli accademici che industriali. Essenzialmente, l'equità nell'IA si riferisce all'assenza di pregiudizi o discriminazioni nei sistemi di IA, un obiettivo complesso a causa delle molteplici forme di pregiudizio che possono manifestarsi. Diversi tipi di equità sono stati proposti in letteratura, tra cui l'equità di gruppo, l'equità individuale e l'equità controfattuale. Sebbene l'equità e il pregiudizio siano concetti strettamente collegati, essi differiscono in modo significativo: l'equità è un obiettivo intenzionale e deliberato, mentre il pregiudizio può essere non intenzionale. Per raggiungere l'equità nell'IA è necessaria una considerazione approfondita del contesto e degli stakeholder coinvolti. Esempi concreti di applicazione dell'equità nell'IA evidenziano i potenziali benefici di incorporare tali principi nei sistemi di IA.

#### Definizione di Equità nell'IA e dei Suoi Diversi Tipi

L'equità nell'IA è un concetto complesso e sfaccettato, oggetto di numerosi dibattiti sia nella comunità accademica che industriale. Fondamentalmente, l'equità si riferisce all'assenza di pregiudizi o discriminazioni nei sistemi di IA. Tuttavia, raggiungere l'equità può essere difficile, poiché richiede un'attenta considerazione dei diversi tipi di pregiudizio che possono sorgere e dei modi per mitigarli.

Esistono vari tipi di equità proposti in letteratura, tra cui:

* **Equità di gruppo**: Garantisce che diversi gruppi siano trattati in modo equo o proporzionale nei sistemi di IA. Questo può includere la parità demografica, che assicura che gli esiti siano distribuiti equamente tra diversi gruppi demografici, la nozione di trattamento disparato, definita in termini di tassi di misclassificazione, o la pari opportunità, che garantisce che il tasso di veri positivi e falsi positivi sia uguale tra diversi gruppi demografici.
* **Equità individuale**: Si riferisce a garantire che individui simili siano trattati in modo simile dai sistemi di IA, indipendentemente dalla loro appartenenza a un gruppo. Questo può essere ottenuto attraverso misure basate sulla similarità o sulla distanza, mirate a garantire che individui con caratteristiche simili siano trattati in modo simile.
* **Equità controfattuale**: Un concetto recente che mira a garantire che i sistemi di IA siano equi anche in scenari ipotetici. In particolare, mira a garantire che un sistema di IA prenderebbe la stessa decisione per un individuo, indipendentemente dalla loro appartenenza a un gruppo, anche se i loro attributi fossero diversi.

Altri tipi di equità includono l'equità procedurale, che comporta garantire che il processo decisionale sia equo e trasparente, e l'equità causale, che comporta garantire che il sistema non perpetui pregiudizi e disuguaglianze storiche.

È importante notare che questi diversi tipi di equità non sono mutuamente esclusivi e possono sovrapporsi nella pratica. Inoltre, diversi tipi di equità possono entrare in conflitto tra loro, richiedendo compromessi per raggiungere l'equità in contesti specifici. Raggiungere l'equità nei sistemi di IA richiede una comprensione sfumata di questi diversi tipi di equità e dei modi in cui possono essere bilanciati e prioritizzati.

#### Confronto tra Equità e Pregiudizio nell'IA

Sebbene l'equità e il pregiudizio siano concetti strettamente correlati, differiscono in modi significativi. Il pregiudizio si riferisce alla deviazione sistematica e coerente dell'output di un algoritmo dal valore reale o da ciò che ci si aspetterebbe in assenza di pregiudizio. L'equità nell'IA, d'altro canto, implica l'assenza di discriminazione o favoritismo verso qualsiasi individuo o gruppo basato su caratteristiche protette come etnia, genere, età o religione.

Una differenza chiave è che il pregiudizio può essere non intenzionale, mentre l'equità è un obiettivo deliberato e intenzionale. Il pregiudizio può sorgere a causa di vari fattori, come dati pregiudiziali o progettazione algoritmica, ma l'equità richiede uno sforzo consapevole per garantire che l'algoritmo non discrimini contro alcun gruppo o individuo. In altre parole, il pregiudizio è un problema tecnico, mentre l'equità è una questione sociale ed etica.

Un'altra differenza è che il pregiudizio può essere sia positivo che negativo, mentre l'equità si preoccupa solo del pregiudizio negativo o della discriminazione. Il pregiudizio positivo si verifica quando un algoritmo favorisce sistematicamente un particolare gruppo o individuo, mentre il pregiudizio negativo si verifica quando l'algoritmo discrimina sistematicamente contro un particolare gruppo o individuo. L'equità mira a prevenire il pregiudizio negativo o la discriminazione verso qualsiasi gruppo o individuo.

Nonostante queste differenze, equità e pregiudizio sono strettamente correlati, e affrontare il pregiudizio è un passo fondamentale per raggiungere l'equità nell'IA. Ad esempio, affrontare il pregiudizio nei dati di addestramento o negli algoritmi può ridurre la probabilità di esiti ingiusti. Tuttavia, è cruciale riconoscere che il pregiudizio non è l'unico fattore che può portare all'iniquità, e raggiungere l'equità può richiedere ulteriori sforzi oltre la mitigazione del pregiudizio.

In generale, comprendere le differenze tra equità e pregiudizio è essenziale per sviluppare strategie efficaci per mitigare il pregiudizio e garantire l'equità nei sistemi di IA. Riconoscendo queste differenze e progettando algoritmi e sistemi che prioritizzino l'equità, possiamo garantire che i sistemi di IA siano utilizzati per avvantaggiare tutti gli individui e i gruppi, senza perpetuare o esacerbare le disuguaglianze sociali ed economiche esistenti.

Tipi di Equità e Descrizione

|  |  |  |
| --- | --- | --- |
| Tipo di Equità | Descrizione | Esempi |
| Equità di Gruppo | Garantisce che diversi gruppi siano trattati in modo uguale o proporzionale nei sistemi di IA. Può essere ulteriormente suddivisa in parità demografica, trattamento disparato o pari opportunità. | 1. Parità demografica: esiti positivi e negativi distribuiti equamente tra i gruppi demografici. 2. Trattamento disparato: definito in termini di tassi di misclassificazione. 3. Pari opportunità: il tasso di veri positivi (sensibilità) e il tasso di falsi positivi (1-specificità) sono uguali tra i diversi gruppi demografici. |
| Equità Individuale | Garantisce che individui simili siano trattati in modo simile dai sistemi di IA, indipendentemente dalla loro appartenenza a un gruppo. Può essere raggiunta attraverso metodi basati sulla similarità o sulla distanza. | Utilizzo di misure basate sulla similarità o sulla distanza per garantire che individui con caratteristiche o attributi simili siano trattati in modo simile dal sistema di IA. |
| Equità Controfattuale | Mira a garantire che i sistemi di IA siano equi anche in scenari ipotetici. In particolare, l'equità controfattuale mira a garantire che un sistema di IA avrebbe preso la stessa decisione per un individuo, indipendentemente dalla loro appartenenza a un gruppo, anche se i loro attributi fossero stati diversi. | Garantire che un sistema di IA prenderebbe la stessa decisione per un individuo, anche se i loro attributi fossero stati diversi. |
| Equità Procedurale | Coinvolge garantire che il processo utilizzato per prendere decisioni sia equo e trasparente. | Implementare un processo decisionale trasparente nei sistemi di IA. |
| Equità Causale | Coinvolge garantire che il sistema non perpetui pregiudizi e disuguaglianze storiche. | Sviluppare sistemi di IA che evitino di perpetuare pregiudizi e disuguaglianze storiche. |

**Strategie di mitigazione già proposte per l’equità nelle AI**

Con il continuo incremento dell'uso dell'intelligenza artificiale (IA), garantire l'equità nelle decisioni di questi sistemi diventa sempre più cruciale. L'impiego dell'IA in settori critici come la sanità, le finanze e il diritto ha il potenziale di influenzare profondamente la vita delle persone. Pertanto, è fondamentale che questi sistemi prendano decisioni eque e prive di pregiudizi. A tal fine, sono stati sviluppati diversi approcci, tra cui l'equità di gruppo e l'equità individuale. Tuttavia, questi metodi presentano limitazioni e sfide, come i compromessi tra vari tipi di equità e la difficoltà di definire l'equità stessa. In questa sezione, analizzeremo le strategie di mitigazione dell'equità nell'IA, includendo gli approcci attuali, le sfide e le aree di ricerca futura. Approfondendo la comprensione di queste strategie, possiamo lavorare verso la creazione di sistemi di IA equi, imparziali ed equitativi per tutti.

#### Panoramica degli Approcci Attuali per Garantire l'Equità nell'IA: Equità di Gruppo e Equità Individuale

Garantire l'equità nell'IA è un campo complesso e in evoluzione, con vari approcci sviluppati per affrontare i diversi aspetti dell'equità. Due approcci chiave emersi sono l'equità di gruppo e l'equità individuale.

L'equità di gruppo si concentra sull'assicurare che i sistemi di IA siano equi verso diversi gruppi di persone, come generi, etnie o etnie. Questo approccio mira a prevenire la discriminazione sistematica contro qualsiasi gruppo, utilizzando tecniche come il campionamento riequilibrato, il pre-processamento o il post-processamento dei dati utilizzati per addestrare il modello di IA. Ad esempio, se un modello di IA è addestrato su dati pregiudiziali verso un particolare gruppo, le tecniche di riequilibrio possono creare un dataset bilanciato, rappresentando equamente ogni gruppo. Altre tecniche, come il pre-processamento o il post-processamento, possono regolare l'output del modello per evitare svantaggi ingiusti per qualsiasi gruppo.

L'equità individuale, invece, si occupa di garantire che i sistemi di IA siano equi verso i singoli individui, indipendentemente dalla loro appartenenza a un gruppo. Questo approccio mira a prevenire decisioni sistematicamente pregiudiziali contro certi individui, utilizzando tecniche come l'equità controfattuale o l'equità causale. Ad esempio, l'equità controfattuale assicura che il modello di IA prenda la stessa decisione per un individuo, indipendentemente da etnia o genere.

Oltre all'equità di gruppo e individuale, altri approcci includono la trasparenza, la responsabilità e la spiegabilità. La trasparenza implica rendere visibile agli utenti il processo decisionale del sistema di IA, mentre la responsabilità coinvolge il rendere responsabili gli sviluppatori per qualsiasi danno causato dal sistema. La spiegabilità si riferisce alla capacità di rendere comprensibili agli utenti le decisioni del sistema di IA.

In generale, garantire l'equità nell'IA è una sfida complessa e continua che richiede un approccio multidisciplinare, coinvolgendo esperti di informatica, diritto, etica e scienze sociali. Sviluppando e implementando vari approcci per garantire l'equità, possiamo lavorare per creare sistemi di IA imparziali, trasparenti e responsabili.

#### Discussione delle Limitazioni e delle Sfide di Questi Approcci

Sebbene questi approcci abbiano mostrato risultati promettenti nel promuovere l'equità nell'IA, non sono privi di limitazioni e sfide. Una significativa limitazione è il potenziale per compromessi tra diversi tipi di equità. Ad esempio, gli approcci di equità di gruppo possono comportare trattamenti iniqui per individui all'interno di un gruppo, mentre gli approcci di equità individuale possono non affrontare i pregiudizi sistemici che colpiscono interi gruppi. Inoltre, può essere difficile determinare quali tipi di equità siano più appropriati per un dato contesto e come bilanciarli adeguatamente.

Un'altra sfida è la difficoltà di definire l'equità stessa. Persone e gruppi diversi possono avere definizioni differenti di equità, che possono cambiare nel tempo. Questo può rendere difficile sviluppare sistemi di IA considerati equi da tutti gli stakeholder.

Molti approcci attuali per garantire l'equità nell'IA si basano su metodi statistici e assunzioni che possono non catturare accuratamente la complessità del comportamento umano e del processo decisionale. Ad esempio, le metriche di equità di gruppo possono non considerare l'intersezionalità o i modi in cui diverse dimensioni dell'identità (come etnia, genere e stato socioeconomico) interagiscono e influenzano gli esiti.

Infine, esistono preoccupazioni riguardo le conseguenze non intenzionali e i risultati dannosi derivanti dai tentativi di garantire l'equità nell'IA. Ad esempio, alcuni ricercatori hanno scoperto che i tentativi di mitigare il pregiudizio negli algoritmi di polizia predittiva possono aumentare le disparità etniche negli arresti.

Approcci per Garantire l'Equità nei Sistemi di IA

|  |  |  |  |
| --- | --- | --- | --- |
| Approccio | Descrizione | Esempi | Limitazioni e Sfide |
| Equità di Gruppo | Garantisce che i sistemi di IA siano equi verso diversi gruppi di persone, come persone di diversi generi, etnie o etnie. Mira a prevenire che il sistema di IA discrimini sistematicamente contro qualsiasi gruppo. Può essere raggiunta attraverso tecniche come il ricampionamento, la pre-elaborazione o la post-elaborazione dei dati. | 1. Tecniche di ricampionamento per creare un set di dati bilanciato. 2. Pre-elaborazione o post-elaborazione per regolare l'output del modello di IA. | 1. Può risultare in un trattamento diseguale degli individui all'interno di un gruppo. 2. Può non affrontare i pregiudizi sistemici che influenzano le caratteristiche individuali. 3. Le metriche di equità di gruppo possono non considerare l'intersezionalità. |
| Equità Individuale | Garantisce che i sistemi di IA siano equi verso gli individui, indipendentemente dalla loro appartenenza a un gruppo. Mira a prevenire che il sistema di IA prenda decisioni sistematicamente pregiudizievoli contro certi individui. Può essere raggiunta attraverso tecniche come l'equità controfattuale o l'equità causale. | 1. Equità controfattuale che assicura la stessa decisione indipendentemente dalla etnia o dal genere. | 1. Può non affrontare i pregiudizi sistemici che influenzano interi gruppi. 2. Difficoltà nel determinare quali tipi di equità sono appropriati per un dato contesto e come bilanciarli. |
| Trasparenza | Coinvolge rendere il processo decisionale del sistema di IA visibile agli utenti. | Rendere comprensibili agli utenti le decisioni e i processi del sistema di IA. | Diverse definizioni di equità tra persone e gruppi e definizioni che cambiano nel tempo. |
| Responsabilità | Coinvolge ritenere responsabili gli sviluppatori del sistema per eventuali danni causati dal sistema. | Sviluppatori ritenuti responsabili per decisioni ingiuste prese dai sistemi di IA. | Determinare la responsabilità e affrontare i potenziali danni. |
| Spiegabilità | Coinvolge rendere le decisioni del sistema di IA comprensibili agli utenti. | Fornire spiegazioni chiare delle decisioni del sistema di IA. | Affrontare la complessità del comportamento umano e del processo decisionale. |
| Intersezionalità | Considera i modi in cui diverse dimensioni dell'identità (come etnia, genere e status socioeconomico) interagiscono e influenzano gli esiti. | Sviluppare sistemi di IA che considerino l'interazione delle diverse dimensioni dell'identità. | Affrontare la complessità dell'intersezionalità e garantire l'equità attraverso molteplici dimensioni dell'identità. |

Comprendendo il tipo di bias esistente, lo sviluppatore può identificare la fase in cui il bias è stato introdotto, permettendogli di progettare una strategia di mitigazione mirata per ciascun tipo di bias. Sono stati proposti metodi esistenti per il pre-processing, rimuovendo il bias dai dati prima della costruzione dell'algoritmo, l'in-processing, costruendo un algoritmo con vincoli legati ai bias, e il post-processing, aggiustando le previsioni di output di un algoritmo.