Abstract

L'obiettivo di questo lavoro è sviluppare un framework innovativo per l'identificazione e la mitigazione dei bias nei modelli linguistici di grandi dimensioni (LLM), con particolare attenzione ai contesti in cui i modelli operano come "scatole nere" e non è possibile intervenire direttamente sulla loro struttura interna. Gli LLM sono potenti strumenti in grado di trasformare numerosi settori, ma la loro natura opaca solleva significative preoccupazioni etiche e sociali, soprattutto quando perpetuano o amplificano bias esistenti.

Il framework proposto si concentra su interventi esterni al modello, progettati per identificare i bias e mitigarli senza la necessità di accedere o modificare direttamente l'architettura interna del LLM. Questo approccio prevede diverse fasi chiave: il pre-processamento dei dati, il monitoraggio e l'analisi del comportamento del modello, e l'implementazione di meccanismi di feedback basati sugli output generati.

Una componente centrale del framework è una rete neurale autoaddestrativa, che opera come strumento di monitoraggio e valutazione continua. Questa rete neurale è addestrata per identificare i bias attraverso l'analisi degli input e degli output del LLM, senza interagire con il modello stesso. Essa valuta costantemente la presenza di bias nei risultati generati e suggerisce interventi correttivi esterni, come la modifica dei dati di input o l'applicazione di tecniche di post-processamento per mitigare gli effetti dei bias rilevati.

Inoltre, il framework integra audit periodici e un sistema di feedback degli utenti, che permette di adattare dinamicamente le strategie di mitigazione ai cambiamenti nei dati e nelle condizioni di utilizzo. Questo approccio garantisce che, anche in assenza di accesso al modello interno, sia possibile mantenere elevati livelli di equità e imparzialità nelle applicazioni basate su LLM.

Il framework enfatizza la collaborazione interdisciplinare tra esperti in etica, ingegneria del software e intelligenza artificiale, per sviluppare metodologie efficaci e adattabili. Questo lavoro contribuisce così alla costruzione di un ecosistema di intelligenza artificiale che, pur operando con modelli "scatola nera", riesce a mitigare in modo proattivo i bias e a promuovere una maggiore equità.

In conclusione, questo lavoro introduce un framework per la mitigazione dei bias negli LLM, basato su interventi esterni e su una rete neurale autoaddestrativa. Anche in contesti in cui il modello è una "scatola nera", il framework offre strumenti potenti per identificare e ridurre i bias, contribuendo a un uso più responsabile e giusto dell'intelligenza artificiale.

(DA RIVEDERE IN CORSO DI STUDIO-BOZZA)

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

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.

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 "AI 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, Definizione e Tipi di Pregiudizio nell’AI

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.

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 riguardante i 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.

Il test dei pregiudizi negli LLM presenta significative sfide ingegneristiche. In primo luogo, richiede competenze provenienti da diversi domini: etica, test software, ingegneria dei prompt e LLM. I test esaustivi richiedono risorse di calcolo significative, portando a considerevoli costi e impronta di carbonio. Inoltre, gli LLM proprietari sono offerti come servizio, con aggiornamenti che possono avvenire silenziosamente e introdurre cambiamenti sottili nelle loro risposte. Potrebbero essere necessari test periodici per garantire che non siano stati introdotti nuovi pregiudizi. Infine, ogni applicazione può essere interessata a diversi tipi di pregiudizi, specifici di un particolare dominio. Di conseguenza, è necessario consentire la definizione delle preoccupazioni etiche e la configurazione del processo di test. L'ingegneria basata su modelli può fornire una soluzione a queste sfide.

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.

Nella formalizzazione matematica dei criteri di equità, gli studiosi si sono confrontati con la complessa questione di quale tipo di uguaglianza debba essere considerata giusta. Alcuni ricercatori sostengono che qualsiasi disparità nelle metriche di risultato sia intrinsecamente inaccettabile. Altri, invece, presuppongono un "campo di gioco livellato", ovvero un contesto in cui non esiste discriminazione di genere o etnica che possa influenzare i dati. La giustizia di una disparità nelle metriche di equità tra gruppi legalmente protetti dipende dal contesto specifico. Ad esempio, etnia e genere possono essere fattori causalmente rilevanti in alcune diagnosi mediche differenziali, come nel caso dell'anemia falciforme o del cancro ovarico, a causa dei diversi meccanismi biologici coinvolti.

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 AI viene utilizzato per confermare pregiudizi o credenze preesistenti dei suoi creatori o utenti. | Un sistema di AI 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 AI 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 AI 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'AI

I sistemi di Intelligenza Artificiale (AI) 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 AI in vari settori industriali, la necessità di comprendere, rilevare e mitigare il bias nelle applicazioni AI ha attirato un notevole interesse. Il bias indesiderato nell'AI 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 AI. È stato osservato che il bias può causare una prestazione subottimale dell'AI 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’AI.

Nella maggior parte delle applicazioni di AI, il bias può derivare da molte fonti. Tra queste fonti comuni di bias nelle applicazioni di AI 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 AI. L'obiettivo e l'implementazione dell'AI 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 AI 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 AI**: 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 AI 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.

# 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 (AI) è solo il primo passo; la sfida successiva riguarda la misurazione di questi bias. In alcuni casi, il bias di un sistema di AI 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'AI, 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 AI, dimostrando che la misurazione del bias dell'AI 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 AI.

### Misurare il Bias dell'AI

Oltre alla misurazione del bias per un sistema di AI, 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 AI 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 AI 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 AI, 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 AI 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'AI.

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

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 AI 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'AI 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 AI 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)

Grazie alla loro crescente diffusione, i grandi modelli linguistici (LLM) come ChatGPT e GPT-4 sono diventati un argomento centrale di dibattito, soprattutto in relazione al problema del bias. Un modello linguistico è un sistema di apprendimento automatico che processa il linguaggio naturale e viene utilizzato per compiti come la risposta a domande, la sintesi e la classificazione del testo. Tuttavia, come qualsiasi prodotto di intelligenza artificiale, anche gli LLM sono suscettibili al bias in ogni fase del loro sviluppo, in particolare durante la raccolta e l'impiego dei dati di addestramento. Sebbene gli sviluppatori di ChatGPT e altri LLM adottino misure per eliminare dati di bassa qualità, offensivi o potenzialmente dannosi, questi modelli possono comunque riflettere bias presenti nei dati su cui sono stati addestrati. Questo rende la rilevazione e la mitigazione del bias una sfida cruciale per le ricerche future. Inoltre, è fondamentale valutare l'autenticità dei dati di addestramento utilizzati, poiché i set di dati pubblici possono essere distorti o contenere "fake news".

Data la loro complessità e le loro dimensioni, i grandi modelli linguistici come ChatGPT e GPT-4 pongono sfide specifiche nella mitigazione del bias. Affrontare efficacemente questo problema richiede uno sforzo collaborativo tra sviluppatori di AI, utenti e le comunità coinvolte.

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 AI più equi e inclusivi, riducendo il bias verso individui o gruppi svantaggiati.

### 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'AI. Questo comporta l'aggiustamento dell'output dei modelli di AI 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'AI.

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'AI 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 AI generativa rimangano equi ed equitativi nel tempo. Questi sforzi devono essere sostenuti da un impegno verso i principi etici dell'AI, coinvolgendo attivamente team diversificati nello sviluppo dell'AI e promuovendo la collaborazione interdisciplinare per affrontare e mitigare efficacemente i pregiudizi nell'AI.

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 AI 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'Intelligenza Artificiale

L'equità nell'Intelligenza Artificiale (AI) è un aspetto cruciale e complesso che ha acquisito crescente attenzione nei contesti accademici e industriali. Questo concetto si riferisce essenzialmente all'assenza di pregiudizi e discriminazioni nei sistemi di AI, un obiettivo intrinsecamente difficile da raggiungere a causa della varietà e della natura dei pregiudizi che possono manifestarsi. I diversi tipi di equità proposti nella letteratura includono equità di gruppo, equità individuale ed equità controfattuale, ciascuno dei quali affronta l'equità da prospettive uniche. Anche se equità e pregiudizio sono concetti strettamente interconnessi, si differenziano significativamente: l'equità è un obiettivo intenzionale e deliberato, mentre il pregiudizio può essere non intenzionale. Per ottenere equità nell'AI è necessaria una considerazione approfondita del contesto e degli stakeholder coinvolti. Gli esempi concreti di applicazione dell'equità nell'AI dimostrano i benefici di integrare tali principi nei sistemi di AI.

L'equità nell'AI è un concetto complesso e sfaccettato, oggetto di vivaci discussioni nella comunità accademica e industriale. Fondamentalmente, essa si riferisce all'assenza di pregiudizi o discriminazioni nei sistemi di AI. Tuttavia, raggiungere l'equità è problematico poiché implica una considerazione accurata dei diversi tipi di pregiudizi che possono emergere e delle modalità per mitigarli. I principali tipi di equità proposti includono:

* **Equità di Gruppo:** Questo tipo di equità si concentra sull'assicurare che diversi gruppi siano trattati in modo equo o proporzionale nei sistemi di AI. Si suddivide in varie sottocategorie:
  + **Parità Demografica:** Garantisce che gli esiti positivi e negativi siano distribuiti equamente tra diversi gruppi demografici. Ad esempio, se un sistema di IA viene utilizzato per determinare prestiti bancari, la parità demografica assicurerebbe che le decisioni di prestito siano equamente distribuite tra diverse etnie e generi.
  + **Trattamento Disparato:** Si riferisce a situazioni in cui i tassi di errore, come la misclassificazione, sono differenti tra gruppi. Per esempio, un algoritmo di riconoscimento facciale potrebbe avere una precisione più alta per un'etnia rispetto ad un'altra, il che rappresenta un trattamento disparato.
  + **Pari Opportunità:** Mira a garantire che il tasso di veri positivi (sensibilità) e il tasso di falsi positivi (1-specificità) siano uguali tra diversi gruppi. Questo implica che un sistema di AI dovrebbe avere la stessa probabilità di identificare correttamente le persone che appartengono a gruppi diversi e di commettere errori.
* **Equità Individuale:** Questa forma di equità si occupa di garantire che individui simili ricevano un trattamento simile, indipendentemente dal gruppo di appartenenza. Può essere realizzata utilizzando misure basate sulla similarità o distanza. Per esempio, un sistema di AI che valuta il rischio di recidiva dovrebbe trattare due individui con lo stesso profilo di rischio in modo equivalente, a prescindere dalla loro etnia o genere.
* **Equità Controfattuale:** Questo concetto relativamente nuovo mira a garantire che i sistemi di AI siano equi anche in scenari ipotetici. In pratica, l’equità controfattuale cerca di assicurare che un sistema di AI avrebbe preso la stessa decisione per un individuo anche se i loro attributi (come etnia o genere) fossero stati diversi. Per esempio, un sistema di selezione del personale dovrebbe dare la stessa raccomandazione a un candidato, indipendentemente dal fatto che il candidato fosse un uomo o una donna, se le loro qualifiche fossero identiche.
* **Equità Procedurale:** Questo tipo di equità si concentra sulla giustizia e sulla trasparenza del processo decisionale. Riguarda la necessità di rendere il processo decisionale chiaro e giustificabile per tutti gli stakeholder coinvolti. Per esempio, i criteri utilizzati da un sistema di AI per approvare un prestito dovrebbero essere pubblicamente accessibili e comprensibili.
* **Equità Causale:** Riguarda la garanzia che il sistema non perpetui pregiudizi e disuguaglianze storiche. Si tratta di sviluppare sistemi di AI che non solo evitano di amplificare le disuguaglianze esistenti, ma che contribuiscano a correggerle. Per esempio, un sistema di AI usato per l'assunzione dovrebbe essere progettato in modo da evitare di replicare o esacerbare le disuguaglianze di genere storicamente esistenti nel campo.

È importante sottolineare che questi diversi tipi di equità non sono mutuamente esclusivi e possono sovrapporsi nella pratica. Inoltre, essi possono entrare in conflitto tra loro, richiedendo compromessi specifici per raggiungere l'equità in contesti particolari. Raggiungere l'equità nei sistemi di IA implica una comprensione dettagliata di questi concetti e un bilanciamento adeguato delle loro applicazioni.

### Confronto tra Equità e Pregiudizio nell'IA

Nonostante la loro interrelazione, equità e pregiudizio presentano differenze significative. Il pregiudizio si riferisce alla deviazione sistematica e coerente dell'output di un algoritmo dal valore reale o atteso in assenza di pregiudizio. Al contrario, l'equità implica l'assenza di discriminazione o favoritismo verso individui o gruppi basati su caratteristiche protette come etnia, genere, età o religione.

Una differenza cruciale è che il pregiudizio può essere non intenzionale, mentre l'equità è un obiettivo deliberato. Il pregiudizio può emergere da vari fattori, come dati distorti o progettazione algoritmica, ma l'equità richiede uno sforzo consapevole per garantire che l'algoritmo non discrimini. In altre parole, il pregiudizio rappresenta un problema tecnico, mentre l'equità è una questione sociale ed etica.

Inoltre, il pregiudizio può essere sia positivo che negativo. Il pregiudizio positivo si verifica quando un algoritmo favorisce sistematicamente un particolare gruppo o individuo, mentre il pregiudizio negativo si manifesta quando l'algoritmo discrimina sistematicamente un gruppo o individuo. L’equità si concentra principalmente sulla prevenzione del pregiudizio negativo, mirando a evitare la discriminazione verso qualsiasi gruppo o individuo.

Nonostante queste differenze, equità e pregiudizio sono strettamente correlati. Affrontare il pregiudizio è un passo fondamentale per raggiungere l'equità nelle AI. Ad esempio, correggere i pregiudizi nei dati di addestramento o negli algoritmi può ridurre la probabilità di esiti ingiusti. Tuttavia, è essenziale riconoscere che il pregiudizio non è l'unico fattore che contribuisce all'iniquità; raggiungere l'equità può richiedere ulteriori sforzi oltre alla semplice mitigazione del pregiudizio.

In generale, comprendere le differenze tra equità e pregiudizio è cruciale per sviluppare strategie efficaci per mitigare il pregiudizio e garantire l'equità nei sistemi di AI. Riconoscere queste differenze e progettare algoritmi e sistemi che prioritizzano l'equità può aiutare a garantire che i sistemi di AI siano utilizzati per beneficiare tutti gli individui e i gruppi, evitando di perpetuare o amplificare le disuguaglianze sociali ed economiche esistenti.

### Approcci per Garantire l'Equità nei Sistemi di IA

#### Panoramica degli Approcci Attuali

Garantire l'equità nei sistemi di AI è un compito complesso e in continua evoluzione. Varie strategie sono state sviluppate per affrontare i diversi aspetti dell'equità, tra cui l'equità di gruppo e l'equità individuale. Ogni approccio ha i suoi metodi e strumenti specifici, ma anche le proprie limitazioni e sfide.

* **Equità di Gruppo:** Questo approccio si concentra sul garantire che i sistemi di AI trattino equamente diversi gruppi di persone, come quelli suddivisi per genere, etnia o altre caratteristiche demografiche. Tecniche comuni per raggiungere l’equità di gruppo includono:
  + **Ricampionamento:** Tecnica che modifica i dati di addestramento per ottenere una rappresentazione più equilibrata tra i gruppi. Ad esempio, se un dataset di addestramento contiene più dati su un gruppo rispetto a un altro, il ricampionamento può essere usato per bilanciare questa rappresentazione.
  + **Pre-Processamento:** Modifica dei dati prima che vengano utilizzati per addestrare il modello, per ridurre i pregiudizi esistenti. Ad esempio, i dati possono essere modificati per eliminare i pregiudizi di genere o etnia.
  + **Post-Processamento:** Modifica dei risultati del modello dopo che è stato addestrato, per garantire che le decisioni finali siano equamente distribuite tra i gruppi. Ad esempio, si possono applicare correzioni ai risultati del modello per garantire che le decisioni di prestito siano eque tra diversi gruppi demografici.
* **Equità Individuale:** Si concentra sul garantire che individui simili ricevano trattamenti simili, indipendentemente dal loro gruppo di appartenenza. Le tecniche comuni includono:
  + **Equità Controfattuale:** Cerca di garantire che la decisione dell'AI per un individuo sarebbe stata la stessa, anche se quell'individuo avesse avuto attributi diversi. Ad esempio, un sistema di assunzione dovrebbe trattare due candidati con lo stesso profilo professionale in modo simile, indipendentemente dal loro genere o etnia.
  + **Misure di Similarità:** Utilizza metodi per confrontare individui basati su attributi pertinenti e garantire trattamenti equi per individui simili. Ad esempio, se due persone hanno lo stesso punteggio di rischio di credito, dovrebbero ricevere lo stesso trattamento da un sistema di prestito.

#### Limitazioni e Sfide

Gli approcci attuali per garantire l'equità nei sistemi di AI presentano numerose limitazioni e sfide:

* **Compromessi tra Tipi di Equità:** Gli approcci di equità di gruppo possono causare trattamenti diseguali tra individui all'interno dello stesso gruppo, mentre gli approcci di equità individuale potrebbero non affrontare le disuguaglianze sistemiche. Ad esempio, bilanciare la parità demografica potrebbe non garantire che i singoli individui ricevano un trattamento equo.
* **Difficoltà di Definizione:** Le definizioni di equità possono variare e cambiare nel tempo, rendendo difficile sviluppare sistemi di AI che siano universalmente considerati equi. Le nozioni di equità possono differire a seconda del contesto culturale, sociale e legale.
* **Metriche Statistiche:** Alcuni metodi per misurare l'equità possono non catturare adeguatamente la complessità del comportamento umano e dell’intersezionalità. Le metriche statistiche possono trascurare come diversi gruppi vivono e interagiscono con il sistema.
* **Conseguenze Non Intenzionali:** Tentativi di mitigare il pregiudizio possono avere effetti contrari. Per esempio, un algoritmo progettato per ridurre il pregiudizio di genere potrebbe finire per amplificare altre forme di discriminazione.

Approcci e Sfide Dettagliati

|  |  |  |  |
| --- | --- | --- | --- |
| Approccio | Descrizione | Esempi | Limitazioni e Sfide |
| Equità di Gruppo | Garantisce che diversi gruppi siano trattati equamente. Include tecniche di ricampionamento, pre-processamento e post-processamento. | Ricampionamento dei dati, modifica dei dati o dell'output del modello. | Potrebbe non affrontare le disuguaglianze tra individui all'interno di un gruppo; non sempre riduce le disuguaglianze sistemiche. |
| Equità Individuale | Assicura che individui simili siano trattati in modo simile. Include tecniche come l’equità controfattuale. | Equità controfattuale per garantire decisioni uniformi. | Difficoltà a mantenere l’equità in presenza di disuguaglianze sistemiche; può richiedere compromessi con l’equità di gruppo. |
| Trasparenza | Rende visibile il processo decisionale del sistema di AI. | Fornire spiegazioni chiare delle decisioni. | Differenze nella definizione di trasparenza e comprensibilità; può non catturare la complessità del processo decisionale. |
| Responsabilità | Rende gli sviluppatori responsabili per danni causati dai sistemi di AI. | Definire responsabilità legale e pratica per le decisioni dell'AI. | Determinare e gestire la responsabilità può essere complesso; può richiedere nuove regolazioni legali e normative. |
| Spiegabilità | Fornisce spiegazioni comprensibili delle decisioni del sistema di AI. | Offrire spiegazioni chiare e accessibili delle decisioni. | La spiegabilità può non sempre riflettere il comportamento complesso degli algoritmi; può non risolvere tutte le problematiche di equità. |
| Intersezionalità | Considera l'interazione tra dimensioni multiple dell'identità. | Sviluppare sistemi che considerano diverse dimensioni dell’identità. | Complessità dell'intersezionalità e difficoltà nel bilanciare le diverse dimensioni; può essere difficile integrare nelle pratiche di sviluppo. |

L’implementazione efficace di questi approcci richiede un lavoro multidisciplinare e una collaborazione tra esperti di informatica, diritto, etica e scienze sociali. Questo sforzo è essenziale per sviluppare sistemi di AI giusti e responsabili, capaci di garantire benefici equi a tutti gli individui e gruppi coinvolti.

### Framework Etici per l'IA: Principi e Applicazioni

I framework etici per l'intelligenza artificiale (AI) forniscono una raccolta di principi progettati per guidare la progettazione e l'implementazione di sistemi AI in modo etico. Questi framework generalmente non stabiliscono una gerarchia tra i principi, ma piuttosto presentano un insieme di valori e linee guida che devono essere rispettati quanto più possibile. Tuttavia, in pratica, può essere necessario bilanciare e ponderare questi principi l'uno contro l'altro. Ad esempio, in un'applicazione medica, il principio della beneficenza (che implica massimizzare i benefici per gli utenti) potrebbe dover essere ponderato rispetto al principio della privacy (che richiede di proteggere i dati personali degli utenti).

Molti dei principi inclusi nei framework etici sono specifici per l'AI e si sono evoluti con l'emergere di queste tecnologie. Principi come il bias del sistema e la tracciabilità sono diventati rilevanti principalmente a causa delle peculiarità dei sistemi AI. Questi principi non sono sempre paragonabili agli approcci tradizionali utilizzati nell'etica medica, dove il principio della beneficenza, ad esempio, è ben definito e ampiamente studiato.

Per applicare efficacemente questi principi, è fondamentale tradurli in requisiti di sistema concreti. Tuttavia, questa traduzione non è una semplice operazione di adattamento; richiede la ricerca e lo sviluppo di strumenti, tecniche e tecnologie specifiche. Ad esempio, i principi etici presentati da Morley et al. spesso non forniscono indicazioni chiare sulle specifiche tecniche necessarie per implementare un sistema informatico etico. Molti di questi requisiti sono formulati a un livello di astrazione che può risultare troppo generale per gli sviluppatori. Alcuni requisiti, come la documentazione delle decisioni di un sistema AI, sono relativamente concreti e praticabili. Tuttavia, concetti come "interpretabilità" rimangono ambigui e difficili da implementare in pratica.

Inoltre, molti problemi etici, come la rimozione del bias da un modello, sono complessi e non facilmente specificabili in termini matematici o algoritmici. L'equità, ad esempio, può avere molte interpretazioni diverse, e non è sempre chiaro quale funzione matematica utilizzare per misurare l'equità in un contesto applicativo specifico. Anche se tali funzioni sono difficili da progettare ex ante, ogni sistema AI definisce implicitamente una funzione di equità attraverso il suo funzionamento, rendendo inevitabile la necessità di mappare il concetto di equità su una funzione computabile, sebbene questo sia estremamente complesso senza un contesto applicativo specifico.

Molti dei framework esistenti non considerano il contesto applicativo e non riflettono le pratiche di sviluppo reali dei sistemi AI, come il processo di tentativi ed errori. Inoltre, spesso si presume implicitamente che il software sia controllato da pochi progettisti, ignorando la realtà che i moderni sistemi software, inclusi i sistemi AI, sono spesso il risultato della collaborazione di molti sviluppatori.

Una questione aperta riguarda se i principi etici dovrebbero essere ottimizzati a livello di singolo principio o a livello dell'intero sistema. Ad esempio, è meglio ottimizzare un componente AI per l'equità mentre si considera l'intero sistema per la beneficenza, o è necessario considerare entrambi i principi simultaneamente? Inoltre, non è chiaro se l'etica possa essere affrontata tramite moduli etici, ossia componenti di un sistema progettati per soddisfare specifici requisiti etici. Questo solleva la questione della composizionalità: un sistema globalmente etico può essere ottenuto combinando componenti che rispettano individualmente i principi etici di un framework?

Infine, i framework etici spesso non considerano i contesti aziendali e le dinamiche dei sistemi socio-tecnici che emergono dopo l'implementazione del sistema AI. Pertanto, è essenziale considerare l'intero processo di sviluppo del sistema AI, come dettagliato nella sezione successiva.

#### Componenti Tipiche dei Framework Etici

1. **Concetti:** Nozioni fondamentali rilevanti per discutere gli aspetti etici.
2. **Principio:** Valori etici (ad esempio, beneficenza, spiegabilità).
3. **Preoccupazione:** Come i principi possono essere minacciati dall'uso e dallo sviluppo dei sistemi AI.
4. **Rimedio:** Strategie, regole e linee guida per affrontare le preoccupazioni etiche.

### Processi e Fasi nella Progettazione di Sistemi Etici per l'IA

I framework etici per l'intelligenza artificiale (AI) spesso presentano principi statici, formulati principalmente come proprietà del sistema AI risultante. Tuttavia, una questione cruciale è comprendere quali fasi del processo di progettazione siano necessarie per sviluppare sistemi etici, considerando che diverse problematiche etiche emergono in fasi diverse dello sviluppo.

Per integrare efficacemente gli aspetti etici, è necessario adottare una visione complessiva che trascenda la semplice modellazione dell'AI. Nella pratica dell'analisi dei dati e dell'apprendimento automatico, l'accento è solitamente posto sulla costruzione del modello e sul processo iterativo di apprendimento. Alcuni approcci delineano un processo suddiviso in diverse fasi, come sviluppo del caso aziendale, comprensione dei dati, modellazione, valutazione e implementazione. Altri approcci propongono varianti che includono fasi specifiche come monitoraggio e documentazione separati.

Un focus sull'epistemologia sottostante organizza il processo in fasi che partono dal dominio del problema e proseguono con creazione, comprensione e pre-elaborazione dei dati, costruzione del modello di AI, e test e valutazione. Combinando questi approcci, si ottiene una segmentazione del processo di progettazione in diverse fasi, che coprono dallo sviluppo del caso aziendale fino al monitoraggio continuo.

Il processo può essere ulteriormente dettagliato per includere vari regimi di test e valutazione, fasi di distribuzione (come utenti di prova, primi adottanti e distribuzione su larga scala) e dismissione. Tuttavia, la segmentazione in fasi cattura la maggior parte dei punti decisionali etici rilevanti durante la progettazione di una nuova applicazione di AI. Ecco un'analisi dettagliata delle fasi:

* **Sviluppo del Modello di Business e del Caso d'Uso:** Questa fase è cruciale per considerare i principi di beneficenza e non maleficenza del sistema, assicurandosi che il sistema apporti benefici e non causi danni.
* **Progettazione del Sistema:** In questa fase, questioni come la partecipazione degli stakeholder e la supervisione umana richiedono attenzione particolare per garantire un'adeguata considerazione dei diritti e delle preoccupazioni degli utenti.
* **Creazione dei Dati:** Focalizzata sulla raccolta etica dei dati, questa fase deve affrontare questioni come l'integrità dei dati e i potenziali bias, garantendo che i dati siano raccolti e gestiti in modo equo e accurato.
* **Pre-elaborazione e Costruzione del Modello:** Sebbene queste fasi si concentrino su aspetti tecnici, sono punti cruciali per garantire l'esplicabilità e l'interpretabilità del modello, nonché per identificare e correggere i bias.
* **Test e Valutazione:** Questa fase è essenziale per verificare l'accuratezza del modello, eseguire test contro possibili attacchi e garantire l'auditabilità. È il momento in cui si valutano le prestazioni del sistema in scenari reali o simulati.
* **Distribuzione e Monitoraggio:** La fase di distribuzione valuta l'impatto a lungo termine del sistema sulla società, sulla democrazia e sull'ambiente. Il monitoraggio continuo è cruciale per identificare e apportare miglioramenti basati sull'uso effettivo del sistema.

Processo di sviluppo delle applicazioni AI

|  |  |  |  |
| --- | --- | --- | --- |
| Processo di sviluppo delle applicazioni AI | Saltz e Dewar | Morley et al. | Prem |
| Sviluppo del business e dei casi d'uso | Comprensione del business | Sviluppo del business e dei casi d'uso |  |
| Progettazione del sistema |  | Progettazione |  |
| Creazione dei dati |  |  | Creazione dei dati |
| Comprensione dei dati | Comprensione dei dati... |  | Comprensione dei dati/ Conoscenza |
| Pre-elaborazione | ... preparazione |  | Pre-elaborazione |
| Addestramento del modello | Modellazione | Costruzione dell'applicazione AI | Sistema AI |
| Test e valutazione | Valutazione | Testing | Test e valutazione |
| Distribuzione | Distribuzione | Distribuzione |  |
| Monitoraggio |  | Monitoraggio |  |

### Integrazione degli Aspetti Etici nel Processo di Progettazione dell'IA

I framework etici per l'intelligenza artificiale (AI) forniscono un insieme di principi che rappresentano le proprietà desiderate del sistema finale. Tuttavia, questi principi sono spesso formulati come indicazioni generali piuttosto che come linee guida specifiche per ogni fase del processo di progettazione dell’AI. Pertanto, è essenziale esplorare come questi principi possano essere tradotti e applicati in modo pratico attraverso le diverse fasi dello sviluppo del sistema.

Per integrare efficacemente gli aspetti etici nella progettazione dell'IA, è necessario un approccio sistematico che vada oltre la semplice modellazione e ottimizzazione del modello di AI. Ecco una panoramica dettagliata del processo di progettazione, suddiviso in fasi chiave, con un focus su come i principi etici possano essere considerati in ciascuna fase:

#### Sviluppo del Modello di Business e del Caso d'Uso: In questa fase iniziale, il focus è sulla definizione degli obiettivi e delle finalità del sistema AI. È essenziale considerare i principi etici di beneficenza e non maleficenza, assicurandosi che il sistema apporti benefici significativi senza causare danni. Questo significa valutare come il sistema AI influenzerà gli utenti e la società, e identificare i potenziali rischi etici associati all'implementazione del sistema.

#### Progettazione del Sistema: Durante la fase di progettazione, si definiscono le caratteristiche e le funzionalità del sistema. È importante considerare la partecipazione degli stakeholder e garantire la supervisione umana, per assicurarsi che il sistema rispetti i diritti degli utenti e le normative vigenti. Questo implica il coinvolgimento di esperti e utenti finali nella progettazione, per assicurare che le loro esigenze e preoccupazioni siano adeguatamente considerate e affrontate.

#### Creazione dei Dati: La raccolta e gestione dei dati sono centrali per lo sviluppo di un sistema AI etico. In questa fase, si devono adottare pratiche etiche per la raccolta dei dati, garantendo la loro integrità e riservatezza. È cruciale evitare la raccolta di dati discriminatori o inaccurati e assicurarsi che i dati siano rappresentativi e privi di bias.

#### Pre-elaborazione e Costruzione del Modello: La pre-elaborazione dei dati e la costruzione del modello sono fasi tecniche che preparano i dati per l'addestramento e il funzionamento del modello di AI. Anche se il focus è principalmente tecnico, è importante considerare l'esplicabilità e l'interpretabilità del modello. Le tecniche di pre-elaborazione dovrebbero mirare a ridurre i bias e migliorare la qualità dei dati, mentre la costruzione del modello deve includere misure per garantire che il modello sia comprensibile e giustificabile.

#### Test e Valutazione: Il testing e la valutazione del modello di AI sono essenziali per verificare la sua accuratezza e prestazioni. In questa fase, si eseguono test per identificare e correggere eventuali problemi, inclusi test di robustezza contro attacchi e vulnerabilità. È anche importante valutare se il sistema rispetta i principi etici stabiliti, come l'auditabilità e la trasparenza.

#### Distribuzione: La distribuzione del sistema AI rappresenta il momento in cui il sistema viene reso operativo e accessibile agli utenti finali. È cruciale monitorare l'impatto del sistema sulla società e garantire che continui a rispettare i principi etici. La distribuzione dovrebbe essere effettuata in modo controllato, ad esempio iniziando con utenti di prova e poi estendendosi a una distribuzione più ampia, per raccogliere feedback e fare aggiustamenti necessari.

#### Monitoraggio: Il monitoraggio continuo del sistema AI è fondamentale per valutare l'impatto a lungo termine e per identificare e risolvere eventuali problemi etici emergenti. Questo include il controllo delle prestazioni del sistema e l'analisi dell'efficacia delle misure etiche implementate. Il monitoraggio deve essere continuo e adattabile per rispondere alle evoluzioni nel contesto applicativo e nelle normative.

### Approcci, Metodi e Strumenti per Affrontare Problemi Etici

Negli ultimi anni, sono state proposte diverse soluzioni per affrontare i problemi etici nell'ambito dell'AI, inclusi elenchi di controllo, standard e tecniche matematiche e informatiche, come la protezione della privacy. Questi approcci variano notevolmente, con alcune proposte molto tecniche e concrete, mentre altre sono linee guida generali che richiedono interpretazione e adattamento. Un'importante analisi condotta da Morley e colleghi ha esaminato oltre 100 proposte di strumenti, metodi e ricerche per risolvere vari problemi etici.

#### Metodologia

L'analisi degli approcci pratici per affrontare i problemi etici nei sistemi di AI si basa su un ampio insieme di riferimenti che include articoli accademici, documenti di standardizzazione, pubblicazioni di consulenza e risorse online. Gli approcci sono stati suddivisi in categorie basate su un processo iterativo di categorizzazione e analisi, che ha incluso articoli che trattano più di un problema etico e proposte di vari tipi.

#### Definizione e Risultati degli Approcci

Gli approcci per affrontare i problemi etici nei sistemi di AI possono essere raggruppati in diverse categorie, che includono:

1. **Sintesi**: Panoramiche che riassumono i problemi etici e forniscono esempi pratici. Questi articoli possono includere studi di caso concreti e raccomandazioni di alto livello.
2. **Nozioni**: Articoli che chiariscono concetti e propongono framework e criteri per facilitare la progettazione etica dell'AI. Questo include elenchi di controllo, dichiarazioni e metriche per misurare aspetti etici.
3. **Procedure**: Approcci che supportano il processo di progettazione etica, inclusi modelli di processo, linee guida e standard. Questi documenti forniscono metodologie e regole generali per garantire la conformità etica.
4. **Codice**: Approcci che riguardano la codifica di sistemi AI etici, inclusi metodi algoritmici, modelli di design, librerie software e assistenti software.
5. **Infrastruttura**: Risorse come set di dati online e comunità di esperti che supportano la creazione di sistemi etici. Questo include risorse per l'apprendimento automatico e spazi per il dibattito e la collaborazione.
6. **Educazione**: Materiale educativo, come tutorial e video, che informa sugli aspetti etici dell'AI.
7. **Valutazioni Ex-Post e Accordi**: Misure adottate dopo lo sviluppo del sistema, come audit e modelli di licenza, per garantire che il sistema continui a rispettare gli standard etici.

Questi approcci forniscono una base solida per affrontare i problemi etici nei sistemi di AI, offrendo strumenti e metodi che possono essere utilizzati in vari stadi del processo di progettazione e implementazione. La categorizzazione e l'analisi di questi approcci aiutano a comprendere le soluzioni disponibili e a identificare le aree in cui ulteriori sviluppi potrebbero essere necessari per garantire che i sistemi di AI siano progettati e gestiti in modo etico.

### **Classificazione per approccio:** La classificazione dei riferimenti è stata effettuata in base agli approcci e ai problemi etici trattati. In generale, la classificazione degli approcci è piuttosto diretta, specialmente per gli approcci algoritmici. Tuttavia, alcuni casi sono più complessi, come quelli che propongono approcci misti. In tali situazioni, è stato deciso di classificare alcuni articoli in più categorie, ad esempio quando propongono sia algoritmi sia metriche, oppure quando si concentrano sulla categoria più chiara. In totale, sono stati identificati 137 approcci distinti. Alcuni lavori sono molto generali o difficili da applicare direttamente ai sistemi di intelligenza artificiale, mentre altri presentano concetti piuttosto semplici. Una significativa proporzione degli approcci (26%) è classificata come metodi algoritmici, il 18% si concentra su approcci concettuali e framework, e il 12% riguarda approcci software. Complessivamente, il codice, che include algoritmi e software, rappresenta il 39% degli approcci.

### **Classificazione per obiettivo etico affrontato:** Gli approcci sono stati classificati in base ai problemi etici che mirano ad affrontare. È emerso che una parte considerevole degli approcci affronta problemi etici a un livello generale, senza focalizzarsi su questioni specifiche. Tuttavia, oltre il 50% degli approcci si concentra su tre problemi concreti: privacy, equità e pregiudizi, e spiegabilità. La responsabilità è trattata da una percentuale minore degli approcci. I restanti problemi etici sono affrontati solo dal 13% degli approcci. Questa classificazione mostra che, nonostante la grande attenzione su problemi specifici come privacy, equità e spiegabilità, esiste una varietà di approcci che si occupano di questioni etiche meno frequenti. La forte enfasi su questi temi indica che sono considerati particolarmente critici nel campo dell'etica dell'intelligenza artificiale.

### **Relazioni tra problemi etici e approcci:** Analizzando la relazione tra i problemi etici e gli approcci proposti, si osserva che questioni come equità, spiegabilità e privacy sono frequentemente trattate con metodi algoritmici e approcci software. Questi strumenti tendono a fornire soluzioni tecniche dettagliate per problemi ben definiti. In contrasto, argomenti più generali e problematiche meno comuni sono spesso trattati con approcci concettuali e modelli di processo. Questa osservazione suggerisce che mentre alcuni problemi etici possono essere affrontati in modo concreto con tecniche specifiche, altri richiedono approcci più astratti e contestuali per essere gestiti in modo efficace.

### **Praticabilità:** Una grande proporzione degli approcci proposti si limita a raccomandazioni generali orientate ai principi etici, offrendo indicazioni senza dettagli operativi su come implementare un sistema AI etico. Questi approcci, pur essendo utili per stabilire principi guida, spesso mancano di applicabilità pratica concreta. Gli algoritmi e i metodi tecnici sono invece più dettagliati e mirano a risolvere problemi specifici come spiegabilità o privacy. Tuttavia, questi approcci richiedono competenze avanzate in informatica e non sempre sono facilmente applicabili senza conoscenze specialistiche. Gli strumenti software e le librerie, che includono moduli già programmati, rappresentano un passo avanti in termini di praticabilità immediata, anche se solo pochi approcci offrono software pronto all'uso. Le metriche sono implementabili e specifiche, e forniscono indicatori per valutare e confrontare i sistemi AI. Le procedure passo-passo, come le checklist e i modelli di processo, sono strumenti pratici che guidano la progettazione e aiutano a prevenire errori noti. Le comunità e le reti di esperti, insieme ai set di dati, costituiscono infrastrutture essenziali per la progettazione etica dei sistemi AI. Gli strumenti ex post, come audit e dichiarazioni, sono utilizzati per valutare e migliorare i sistemi già esistenti, offrendo trasparenza e opportunità di miglioramento continuo.

### **Approcci ex ante versus ex post:** Gli strumenti per affrontare le questioni etiche differiscono notevolmente riguardo al momento del loro intervento nel processo di progettazione. I framework, gli algoritmi e le librerie software mirano a influenzare la creazione di sistemi AI etici fin dall'inizio del processo. Al contrario, strumenti come audit, checklist e metriche sono applicati una volta che il sistema è stato sviluppato, con l'obiettivo di migliorare l'eticità del sistema in modo iterativo. Le informazioni, come dichiarazioni e etichette, possono essere applicate ex post, fornendo trasparenza e guida agli utenti e agli organismi di regolamentazione anche quando un sistema etico completo non è realizzabile. La relazione tra gli approcci e le fasi del processo di sviluppo del sistema AI è complessa. Alcuni approcci possono essere utili in più fasi del processo, ma è difficile mappare esattamente il loro impatto. Alcuni strumenti possono avere un effetto che si manifesta principalmente dopo la distribuzione del sistema, come la documentazione dei set di dati o le carte modello per descrivere le caratteristiche del modello addestrato.

### Discussione e Questioni etiche mancanti

L'analisi degli approcci evidenzia un grande interesse per il miglioramento della progettazione di sistemi AI etici e una varietà di proposte da ricercatori e professionisti. C'è una crescente attenzione a problemi specifici come la privacy, l'equità e la spiegabilità, mentre altre questioni etiche ricevono meno attenzione. Ad esempio, il campo del machine learning rispettoso della privacy è ora una sotto-disciplina consolidata, e la spiegabilità è un tema importante nella ricerca sull'AI. I modelli di processo standard, come IEEE P7000, sono tra i primi a formalizzare l'ingegneria dei sistemi etici.

Il fatto che equità e spiegabilità siano spesso affrontate con approcci tecnici suggerisce che questi problemi sono percepiti come caratteristiche intrinseche dei sistemi AI. Tuttavia, affrontare questioni come l'equità a livello sociale e applicativo potrebbe richiedere approcci più complessi, come framework regolamentari o modelli di processo che considerano esplicitamente valori e fattori contestuali.

Nonostante i significativi progressi, alcune questioni etiche rimangono sottorappresentate. Problemi come il controllo democratico, la governance, le minacce esistenziali e i costi sociali e ambientali dell'AI sono raramente trattati. Progettare sistemi con conseguenze gravi non è del tutto nuovo nel campo dell'informatica, e l'ingegneria del software per sistemi critici per la sicurezza ha sviluppato tecniche per gestire fallimenti catastrofici. Anche se esistono lavori iniziali sul controllo democratico dei sistemi, molto resta da fare per sviluppare concetti e framework adeguati.