**INTRODUZIONE**

L'intelligenza artificiale (IA), 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 la scienza e la società nel loro insieme. 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 svariati settori, tra cui la sanità, i trasporti e la produzione industriale.

L'impatto trasformativo dell'IA e le sue implicazioni sociali hanno stimolato un vivace dibattito riguardo ai principi etici e ai valori che dovrebbero orientarne lo sviluppo e l'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, che potrebbe compromettere l'equità. Questi temi sono stati ampiamente discussi sia nella letteratura scientifica recente che nei media.

In risposta a tali preoccupazioni, numerose organizzazioni nazionali e internazionali hanno istituito comitati di esperti dedicati all'IA, 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'IA dell'Organizzazione per la Cooperazione e lo Sviluppo Economico (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'IA. Parallelamente, 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'IA. Anche 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'IA, accompagnata da un significativo aumento della produzione scientifica e delle dimensioni del mercato.

Dall’analisi del contenuto di diversi 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.

**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 IA. La trasparenza è vista principalmente come un mezzo per minimizzare i danni e migliorare l'IA, 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 IA. Le specifiche su cosa debba essere comunicato variano: uso dell'IA, codice sorgente, uso dei dati, base di evidenze per l'uso dell'IA, 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'IA. Gli emittenti del settore pubblico sottolineano l'impatto dell'IA 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'IA 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'IA 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'IA.

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

In questo contesto, il nostro focus si concentra sul principio di equità e imparzialità in relazione ai bias prodotti dai modelli linguistici di grandi dimensioni (LLM); esiste una crescente preoccupazione che i contenuti generati dagli algoritmi di machine learning (ML) possano discriminare ingiustamente basandosi su attributi demografici personali.

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 e mirati per le valutazioni degli algoritmi, risultando difficili da integrare in una valutazione etica più ampia di un algoritmo.

A causa della difficoltà di quantificare i bias rilevanti e di districarli dal risultato di interesse, correggere un bias comporta il rischio di aumentare le imprecisioni delle previsioni.

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

**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.

Abbiamo quindi sottolineato alcuni dei collegamenti filosofici agli algoritmi di machine learning così da far evincere l’importanza che i modelli di IA hanno nella vita odierna, e la fondamentale importanza dell’avere un modello il più generico possibile che non sia soggetto a bias.

Per far 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.