Digeat N.1 - 19 Marzo 2024

Affrontare i bias di genere nell’IA: percorsi per un’equità tecnologica

Di Adriana Augenti

Abstract

I pregiudizi possono essere una realtà ineludibile nella vita quotidiana che incide sul modo in cui assimiliamo le informazioni ogni giorno: la distorsione dei dati. Ma esiste la possibilità di evitare che diventino una componente delle tecnologie emergenti? Il funzionamento dei Modelli di Large Language Models (LLMs) riflette il sapere condiviso della società, emergendo da contesti sia espliciti che impliciti presenti nei dati usati per l’addestramento. L’evoluzione dell’intelligenza artificiale (IA) può (e sta) trasformando velocemente l’ambiente lavorativo. Questi sviluppi offrono la possibilità di promuovere la parità di genere, ma rischiano anche di perpetuare stereotipi, sessismo e discriminazione nel mondo del lavoro. Il problema del *bias* nei dati nella selezione del personale è influenzato non solo dagli stereotipi e pregiudizi incorporati nei meccanismi di addestramento dell’IA, ma anche dal fatto concreto che le donne occupano in numero minore molte professioni, incidendo sulla neutralità dei sistemi basati sull’intelligenza artificiale. Lo sviluppo di un modello di linguaggio del tutto equo comporta notevoli difficoltà, ma, adottando approcci olistici che includano tecnologie avanzate, norme etiche e il contributo di una vasta gamma di prospettive, è possibile fare passi avanti verso la realizzazione di sistemi di IA più giusti e imparziali.

Indice

- I mercati del lavoro
- Gender Gap e Intelligenza Artificiale
- Il linguaggio
- Verso un’equità tecnologica
- L’equità assoluta rimane un obiettivo sfidante!

I mercati del lavoro

La tecnologia da sempre agisce come un potente catalizzatore per la crescita della produttività, rappresentando uno dei pilastri fondamentali del progresso umano. Le innovazioni tecnologiche sono in grado di trasformare radicalmente il modo in cui lavoriamo, rendendo i processi più efficienti, riducendo il dispendio di tempo e di risorse, e sbloccando nuove possibilità che prima erano inimmaginabili.Il progresso tecnologico non si limita solamente a un’ottica di efficienza e crescita economica, ma si estende al benessere dei lavoratori e al miglioramento delle condizioni di vita, incarnando i valori di giustizia, equità e dignità umana.

> Nella sua essenza, la tecnologia rappresenta una forza di emancipazione e progresso, che, se indirizzata e disciplinata saggiamente, può contribuire significativamente a costruire società più giuste, inclusive e sostenibili. L’adozione e

> l'applicazione delle tecnologie algoritmiche, come l'intelligenza artificiale e l'apprendimento automatico, promettono rivoluzioni nel modo in cui viviamo, lavoriamo e interagiamo. Tuttavia, senza un impegno consapevole e proattivo, questi strumenti possono involontariamente perpetuare e persino amplificare le disuguaglianze sociali esistenti e rafforzare stereotipi dannosi. Questo fenomeno si verifica perché gli algoritmi, nonostante la percezione di neutralità e obiettività, sono creati e addestrati da esseri umani e, pertanto, orientati ad ereditare i pregiudizi impliciti e le visioni del mondo dei loro creatori.

Un recente studio condotto dal FMI ha rivelato che circa **il 40% dell'occupazione a livello mondiale risulta vulnerabile all'Intelligenza Artificiale**, evidenziando una esposizione anche maggiore nei paesi con economie avanzate. Al contempo, gli algoritmi avanzati stanno sfidando le aspettative, dimostrando la loro capacità di affiancare e persino rimpiazzare posizioni altamente specializzate che si ritenevano al riparo dall'automazione.

Questo fenomeno sta **ridefinendo il tessuto dei mercati del lavoro e alterando le previsioni sul futuro dell'occupazione**, sfumando la netta distinzione tra lavori a rischio di automazione e quelli ritenuti sicuri, mostrando come l'avanzamento tecnologico possa interessare l'intero spettro professionale.
In questo contesto, le donne, fortemente rappresentate nel settore dei servizi, e le lavoratrici e i lavoratori con un elevato livello di istruzione, tipicamente impegnati in professioni che necessitano di un'intensa attività cognitiva, si trovano in una **posizione di maggiore vulnerabilità**. Nonostante ciò, entrambi i gruppi possiedono significative opportunità di beneficiare dall'integrazione dell'IA, avendo così la possibilità di capitalizzare i vantaggi che questa tecnologia avanzata può offrire.

Gender Gap e Intelligenza Artificiale

È fondamentale sottolineare che le donne affrontano rischi specifici legati alla propensione al *bias* negli algoritmi di intelligenza artificiale.

> Nell'ambiente tecnologico, la discriminazione di genere causata dagli algoritmi può manifestarsi in modi sottili, ma significativi, influenzando non solo le carriere ma anche, ad esempio, come i prodotti, i servizi e le offerte vengono sviluppati e a chi si rivolgono.

L'uso non critico di tecnologie algoritmiche rischia di amplificare stereotipi dannosi.
Ad esempio, gli algoritmi di raccomandazione possono **intrappolare gli utenti in bolle informative** che rafforzano pregiudizi e visioni del mondo limitate, piuttosto che esporli a una varietà di prospettive. Gli **algoritmi di raccomandazione** usati dai servizi di streaming video possono perpetuare stereotipi di genere suggerendo contenuti basati su dati storici che riflettono pregiudizi, come l'associazione di certi generi cinematografici a specifici generi sessuali. Questa discriminazione **limita l'esposizione delle donne a una varietà di contenuti e potrebbe escluderle da quelli di loro interesse**, influenzando negativamente sia le loro esperienze come consumatrici sia le opportunità come creatrici nel campo tecnologico. La progettazione e il marketing di prodotti tecnologici basati su stereotipi di genere possono ulteriormente esacerbare la disparità di genere, mentre il bias nei dati utilizzati per addestrare l'IA può creare un **ciclo di discriminazione che si autoalimenta.**

Google ha affrontato problemi specifici legati alla discriminazione di genere nella sua funzione di ricerca di lavoro, dove è stato riscontrato che gli algoritmi tendevano a mostrare annunci per posizioni lavorative di alto livello più frequentemente agli uomini piuttosto che alle donne. Questo comportamento degli algoritmi riflette una forma di bias algoritmico, dove, a causa di pregiudizi presenti nei dati di addestramento o nelle assunzioni di sviluppo, **le donne venivano meno esposte a opportunità lavorative di rilievo rispetto ai loro colleghi uomini.**Se gli algoritmi favoriscono gli uomini nella presentazione di annunci per ruoli di leadership o altamente retribuiti, ciò non solo limita

l'accesso delle donne a queste opportunità, ma contribuisce anche a mantenere le barriere strutturali che impediscono l'uguaglianza di genere nel mondo del lavoro.

Un altro caso di discriminazione di genere è rappresentato dal ***deployment bias* nell'assegnazione di credito finanziario**. Questo emerge come un problema significativo, particolarmente nei confronti delle donne imprenditrici. A causa della dipendenza da dati storici che riflettono l'accesso limitato al credito e i pregiudizi di genere, un modello di machine learning sviluppato da una banca per valutare le richieste di prestito può perpetuare la discriminazione, valutando le imprenditrici come più rischiose, non solo limitando ingiustamente l'accesso delle donne ai finanziamenti necessari per le loro attività, ma minando anche la diversità e l'innovazione nell'ecosistema imprenditoriale.
Il **caso della Apple Card**, introdotta da Apple in partnership con Goldman Sachs, fornisce un esempio concreto e attinente di come il ***bias* algoritmico possa influenzare l'accesso ai prodotti finanziari**, evidenziando situazioni in cui donne ricevevano limiti di credito inferiori rispetto agli uomini, anche in casi dove le donne avevano punteggi di credito comparabili o addirittura superiori rispetto a quelli maschili.
Sebbene l'indagine sulle accuse di discriminazione di genere legate alla Apple Card non abbia evidenziato discriminazioni illegali secondo le leggi statali (USA) sul *fair lending*, questo episodio ha messo in luce rischi significativi associati all'uso di algoritmi e processi decisionali automatizzati nel settore finanziario. Questi rischi includono la possibilità che tali sistemi possano i**nvolontariamente perpetuare o addirittura amplificare bias esistenti, come quelli di genere**. È necessario valutare accuratamente l'impiego dei modelli di apprendimento automatico in situazioni concrete, soprattutto quando questi influiscono sull'accesso a risorse fondamentali come il credito.

L'uso di algoritmi di *machine learning* nella valutazione delle performance e nella determinazione delle promozioni, nonché nelle assunzioni (**caso Amazon**) o nelle determinazioni salariali, può portare a discriminazione di genere nel posto di lavoro. **Se un algoritmo analizza dati storici in cui gli uomini sono stati promossi più di frequente delle donne, potrebbe erroneamente concludere che il genere maschile è un indicatore di idoneità per la promozione**. Questo accade non perché l'algoritmo sia stato esplicitamente programmato per considerare il genere come fattore, ma perché apprende dai pregiudizi intrinseci nei dati di addestramento. Di conseguenza, tale algoritmo potrebbe ingiustamente favorire i candidati uomini per le promozioni, per le assunzioni, per le determinazioni salariali, perpetuando la discriminazione di genere esistente all'interno dell'azienda.

La riservatezza con cui le aziende gestiscono i propri dati e sistemi automatizzati spesso ostacola la rivelazione pubblica di casi di discriminazione, rendendo complesso identificare episodi specifici senza un'indagine approfondita o la divulgazione attraverso procedure legali. Ciò avviene anche a causa della mancanza di trasparenza e di standardizzazione nel reporting di tali bias, che si traduce in un problema significativo nel campo dell'intelligenza artificiale e della tecnologia: la necessità di maggiore apertura e responsabilità per garantire l'equità e l'inclusione nelle pratiche lavorative automatizzate.

Al fine di contrastare il bias algoritmico è, dunque, **essenziale un approccio multisfaccettato che includa la revisione critica dei set di dati utilizzati per l'addestramento degli algoritmi**, assicurandosi che siano rappresentativi e privi di pregiudizi; l'implementazione di tecniche di apprendimento automatico che identificano e correggono attivamente i bias; e un controllo umano costante delle decisioni prese dagli algoritmi, soprattutto in contesti ad alto impatto come quelle fin qui analizzate.

Il linguaggio

È fondamentale chiarire l'erronea convinzione secondo cui i pregiudizi derivano esclusivamente dall'analisi dei dati. In realtà, tali pregiudizi sono **il frutto di un complesso intreccio di interazioni tra società e tecnologia, che agiscono congiuntamente nella perpetuazione di pregiudizi discriminatori**. Questa dinamica sottolinea come le scelte di design, le norme culturali e le pratiche

sociali si riflettano e si rafforzino attraverso l'uso delle tecnologie, potendo contribuire attivamente alla costruzione e al mantenimento dei *bias*.
Tra queste particolare attenzione merita il linguaggio, che non è solo uno strumento di comunicazione ma anche un mezzo attraverso il quale vengono costruite e perpetuate le percezioni sociali.

Gli assistenti vocali intelligenti come **Siri** e **Alexa** tradizionalmente sono stati programmati con **voci femminili** *di default*. Questi assistenti hanno ricevuto critiche per le loro risposte a commenti o comandi sessisti, in alcuni casi rispondendo in modo sottomesso o flirtante a insulti o avances sessuali. Tale comportamento deriva da decisioni di design che riflettono e perpetuano stereotipi di genere, suggerendo che le figure "femminili" virtuali debbano essere accomodanti o divertenti di fronte a comportamenti sessisti.
L'analisi e l'uso consapevole del linguaggio di genere, inoltre, sono cruciali per diversi motivi, specialmente nel contesto della comunicazione e della **creazione di tecnologie inclusive**: consente di riflettere accuratamente le **diverse identità ed esperienze**, promuovendo l'inclusività; aiuta a contrastare gli stereotipi di genere radicati nella società; influisce sulla percezione del messaggio da parte del destinatario; educa e sensibilizza sulle questioni di genere, promuovendo una maggiore consapevolezza delle discriminazioni e delle disparità.

Nello sviluppo di prodotti tecnologici, come software, applicazioni e assistenti virtuali, **l'uso del linguaggio di genere gioca un ruolo significativo** nel definire come gli utenti interagiscono con queste tecnologie.I modelli di intelligenza artificiale sviluppati per generare immagini a partire da descrizioni testuali illustrano perfettamente l'importanza del linguaggio di genere nello sviluppo di prodotti tecnologici e il suo impatto sull'interazione degli utenti con queste tecnologie: quando gli viene chiesto di creare immagini rappresentative di professioni con una descrizione di genere neutro, il modello tende a produrre immagini che rappresentano prevalentemente uomini, riflettendo così i *bias* di genere presenti nei dati su cui è stato addestrato.
Questo comportamento evidenzia come i preconcetti incorporati nei set di dati possano influenzare direttamente le tecnologie basate sull'intelligenza artificiale, **portando a risultati che perpetuano stereotipi di genere**. Nonostante la neutralità apparente delle richieste, l'algoritmo prosegue nel rafforzare l'idea che certe professioni siano intrinsecamente maschili, limitando la rappresentazione delle donne e di altre identità di genere nelle immagini generate.

> L'esigenza di un linguaggio inclusivo e di una riflessione critica sulle fonti dei dati diventa quindi fondamentale per garantire che le tecnologie siano veramente accessibili e accoglienti per tutti. Migliorare l'esperienza dell'utente attraverso un linguaggio inclusivo significa anche interrogarsi su come gli algoritmi interpretano ed eseguono le richieste, assicurandosi che i prodotti tecnologici riflettano una società diversificata e inclusiva.

In Italia sarà a breve lanciato il progetto E-Mimic, vincitore del bando Prin 2022, sviluppato da un consorzio di università con l'obiettivo di eliminare stereotipi discriminatori dai testi amministrativi e universitari. Utilizzando un algoritmo avanzato, E-Mimic modifica i termini potenzialmente pericolosi nei testi per promuovere una narrazione inclusiva, affrontando la discriminazione di genere, l'età (ageismo), e le disabilità. Un'iniziativa che mira a ispirare riflessioni e cambiamenti positivi nell'uso del linguaggio.

Verso un'equità tecnologica

Al di là dell'analisi dei dati, altri fattori critici concorrono al *bias* algoritmico: le scelte di design e sviluppo; le assunzioni implicite degli sviluppatori; le norme culturali e sociali che si riflettono nei processi decisionali tecnologici. Oltre all'addestramento con dati storici, **il *bias* può insinuarsi anche in momenti successivi, come nella selezione delle features, nell'interpretazione dei risultati e nell'interfaccia utente**, che può involontariamente guidare verso interazioni viziate. L'equità

tecnologica richiede un impegno consapevole e proattivo che includa una varietà di prospettive e competenze nella creazione e valutazione degli algoritmi.

L’equità assoluta rimane un obiettivo sfidante!

L’intelligenza artificiale (IA) ha un impatto mondiale e la frammentazione legislativa potrebbe limitare lo sviluppo tecnologico e, paradossalmente, perpetuare il bias a causa delle stesse diversità culturali e socioeconomiche che le diverse normative mirano a proteggere.

Al fine di garantire **uno sviluppo etico e responsabile delle tecnologie IA** bisognerebbe: implementare standard internazionali di trasparenza, per assicurare che l’uso dei dati e le decisioni algoritmiche siano chiari e che gli sviluppatori rendano conto dei potenziali bias; costituire team di sviluppo diversificati, che riflettano la varietà della popolazione generale, per ridurre il rischio di pregiudizi inconsci e favorire l’equità nelle tecnologie sviluppate; effettuare audit esterni regolari, per controllare l’imparzialità dei sistemi IA e verificarne la conformità ai principi di equità; elaborare misure specifiche per la protezione dei dati personali degli utenti; promuovere la collaborazione internazionale, per scambiare conoscenze e risorse utili a combattere il bias algoritmico, condividendo pratiche efficaci e innovazioni.

Incorporando questi principi nel cuore delle politiche pubbliche e delle strategie aziendali, può essere possibile avanzare **verso un futuro tecnologico che sia non solo innovativo, ma anche inclusivo ed etico**.

L’**AI Act**, recentemente approvato dal Parlamento Europeo, rappresenta un passo avanti significativo nella regolamentazione dei sistemi di intelligenza artificiale, introducendo un approccio graduale ai rischi che varia da inaccettabili a minimi. Questa normativa impone ai fornitori di IA di **adottare misure di trasparenza**, consentendo agli utenti di comprendere il funzionamento degli algoritmi e i dati sottostanti. In particolare, i sistemi ad alto rischio devono affrontare valutazioni di conformità prima di essere introdotti sul mercato, assicurando l’aderenza ai principi di sicurezza, non discriminazione e rispetto dei diritti fondamentali.

L’AI Act stabilisce anche u**n quadro per la sorveglianza e la supervisione post-lancio dei sistemi IA**, per monitorarne la sicurezza e la conformità nel tempo. Un elemento chiave della legislazione è **l’enfasi sulla necessità di meccanismi di rimedio efficaci**, che permettano di affrontare e correggere rapidamente qualsiasi violazione dei principi etici e normativi, inclusi quelli relativi al *bias* algoritmico.

In attesa di verificare l’efficacia dell’AI Act, probabilmente se un approccio simile fosse adottato su scala globale, si potrebbero **stabilire standard comuni che rispettino le diversità culturali e socioeconomiche**, promuovendo lo sviluppo di un’IA che possa essere sia innovativa sia eticamente responsabile.