



Rassegna Stampa

Agenda Digitale

DISTORSIONI ALGORITMICHE

AI bias: come i pregiudizi influenzano l'intelligenza artificiale

Il bias nell'IA rappresenta distorsioni sistematiche nei risultati algoritmici, influenzati da dati di addestramento e pregiudizi umani. Questo fenomeno può portare a discriminazioni e disinformazione, richiedendo soluzioni per un'IA più etica e inclusiva

Pubblicato il 12 mar 2025

Juliette Blackman

Ughi e Nunziante Studio Legale

Emanuela Burgio

Ughi e Nunziante Studio Legale



I termine inglese "*bias*" deriva dal francese "*biais*", che significa letteralmente obliquo o inclinato. **Nell'ambito dell'intelligenza artificiale (IA), "*bias*" può avere diverse connotazioni.** Nel campo delle scienze algoritmiche si riferisce a una differenziazione necessaria parametrata per ottenere un determinato risultato: se un algoritmo deve distinguere una mela da un'arancia, dovrà avere un *bias*. Tuttavia, in un'accezione più comune, il *bias* IA indica una distorsione o un pregiudizio che può influenzare i risultati generati dagli algoritmi in modo sistematico e involontario^[1].

Indice degli argomenti

Origini e meccanismi dei pregiudizi algoritmici

Tali *bias* possono derivare da vari fattori: dalle opinioni e inclinazioni dei programmatori, dai dati di addestramento o dalla struttura stessa degli algoritmi che spesso si autocorreggono, rendendo sempre più complesso il monitoraggio. Anche l'*input* – ovvero la domanda formulata dall'utente – riflette inevitabilmente i pregiudizi di chi lo formula. Tutti questi passaggi possono contaminare l'*output* finale, con conseguenze tutt'altro che trascurabili, tra cui la **diffusione di stereotipi, discriminazioni e disinformazione**.

Nonostante lo stesso ChatGPT ammetta l'esistenza di tali pregiudizi, molti continuano a credere erroneamente che la tecnologia sia neutra e obiettiva[2]. Questo articolo mira a offrire una prospettiva critica illustrando casi concreti di *bias* e presentando delle soluzioni in fase di sviluppo al fine promuovere un'IA più etica.

Esempi concreti di pregiudizi nei sistemi di IA

Un primo esempio di bias è quello dei sistemi di IA per l'**identificazione biometrica**, che in tempo reale riescono a riconoscere delle caratteristiche umane fisiche, comportamentali o psicologiche allo scopo di determinare l'identità di una persona confrontando i suoi dati biometrici con quelli di altri individui memorizzati in una banca dati, indipendentemente dal fatto che la persona abbia o meno fornito il proprio consenso[3]. Questi sistemi di identificazione biometrica sono ampiamente utilizzati dalle forze di polizia che si avvalgono di database contenenti foto segnaletiche e patenti di guida[4].

È stato notato, però, come questi programmi, siano **meno precisi con le donne e le persone di determinate etnie**, a causa della carenza di diversità nei dati di addestramento, costituiti prevalentemente da immagini di uomini bianchi tra i 18 e i 45 anni[5]. Questi esempi illustrano un chiaro **data bias**, che riproduce nel sistema di intelligenza artificiale tendenze discriminatorie, stereotipi o disuguaglianze già presenti nella società.

“Pensare che oggi l'intelligenza artificiale non possa replicare la nostra storia è un po' un'illusione: è diventato un sistema discriminatorio semplicemente perché è lo specchio della nostra società.”[6]

Giustizia predittiva e bias razziale

A tal proposito, si noti come la giustizia americana ricorra sempre più all'intelligenza artificiale

nelle varie fasi del processo penale, soprattutto per formulare **giudizi predittivi sul rischio di criminalità o recidiva**. Questi sistemi analizzano vari dati quali il contesto familiare, il tasso di criminalità del quartiere, l'età, l'occupazione, i precedenti penali e altri fattori utili per stimare tale rischio[7]. Uno degli algoritmi più noti in questo ambito è COMPAS[8], che genera punteggi del rischio di recidiva e violenza attraverso un questionario composto da 137 domande[9].

Tuttavia, COMPAS ha mostrato evidenti segni di **racial bias**: gli imputati neri hanno quasi il doppio delle probabilità di essere erroneamente classificati come ad alto rischio rispetto ai bianchi (il 45% contro il 23%). Al contempo, un'elevata percentuale di bianchi con recidiva sono stati valutati a basso rischio[10]. Per di più, la scarsa trasparenza dell'algoritmo, il cui funzionamento è segreto e brevettato, ha sollevato questioni sul diritto al giusto processo, che tuttavia sono state respinte dalla Corte Suprema del Wisconsin. La Corte ha in effetti stabilito che il sistema fosse sufficientemente attendibile poiché il questionario è riportato nel manuale d'uso del *software* e si basa sulle risposte dell'imputato. Pertanto, l'imputato ha la possibilità di verificare l'accuratezza delle risposte riportate nel rapporto COMPAS[11]. La Corte ha, tuttavia, sottolineato che, sebbene COMPAS possa essere utilizzato in fase di condanna, il suo impiego deve essere opportunamente circoscritto[12].

Istruzione, la discriminazione dell'IA nei processi di selezione

Spostando l'indagine al contesto educativo, è stato osservato come negli ultimi anni, università, college e altri istituti di formazione hanno introdotto sistemi di IA nei loro **processi di ammissione** per accelerarli e per poter valutare un maggior numero di candidati. Anche questo settore si è rivelato non immune ai *bias*, e alcuni studenti si sono trovati ad essere svantaggiati per il sol fatto che i dati non hanno rappresentato adeguatamente l'intera popolazione.

Ad esempio, **il sistema GRADE, usato dall'Università del Texas** fino al 2020 per valutare i candidati al dottorato di ricerca è stato addestrato con i dati delle ammissioni degli anni precedenti, analizzando le caratteristiche delle candidature accettate e di quelle respinte. Sono così emersi elementi che aumentano la probabilità di accettazione, come lettere di raccomandazione contenenti le parole "migliore" o "premio", nonché curriculum con voti alti o nomi di università di *élite*. Questo sistema, tuttavia, non rappresenta adeguatamente individui privi delle risorse necessarie per frequentare istituti di *élite*[13].

La discriminazione che ne deriva è particolarmente grave, poiché determina il percorso educativo e professionale di una persona. Di fronte a queste preoccupazioni, l'Università del

Texas ha scelto di abbandonare l'uso di GRADE nel 2020.

Bias di genere (e non solo) nel settore dell'occupazione

Guardando al contesto lavorativo, sono sempre più frequenti i casi di adozione di **sistemi di IA per automatizzare i processi di assunzione del personale**, l'attribuzione di compiti o promozioni e i licenziamenti. I *bias* hanno in questo settore un impatto significativo in termini di prospettive di carriera, sostentamento e diritti dei lavoratori[14].

Un esempio noto è il sistema di selezione del personale di una azienda attiva nel settore della logistica, che penalizzava sistematicamente le **candidature femminili** per il ruolo di sviluppatore di software e altri posti ad alto contenuto tecnologico.

Il bias si era creato perché il sistema era stato addestrato su un database di curricula raccolti dalla stessa società nell'arco di dieci anni e la maggior parte di essi proveniva prevalentemente da uomini. Di conseguenza, il sistema ha auto-dedotto che i candidati maschili fossero più idonei, penalizzando i curricula con riferimenti a esperienze o qualifiche femminili. Riflettendo uno squilibrio di genere preesistente nel settore tecnologico, l'azienda ha impiegato anni per identificare e affrontare il problema[15].

Le cause del bias nell'intelligenza artificiale

Analizzando la casistica sopra proposta, ci si potrebbe chiedere se l'**IA** sia ontologicamente discriminatoria. Evidentemente, poiché si sta parlando di una macchina, si può escludere che il pregiudizio dell'IA non sia un "pensiero" della stessa. Di contro, il problema dei bias è intrinsecamente umano.

L'IA apprende analizzando grandi quantità di dati storici e individuando schemi che poi applica nelle sue previsioni o decisioni future. Tuttavia, se i dati riflettono disuguaglianze, stereotipi o pregiudizi esistenti nella realtà sociale, economica o culturale, il sistema tenderà a replicarli. Ad esempio, se un modello viene addestrato su dati storici di assunzioni in cui le donne sono state sistematicamente sottorappresentate in posizioni dirigenziali, l'IA potrebbe inconsapevolmente considerare la presenza femminile in tali ruoli come anomala o meno desiderabile, perpetuando un'ingiustizia storica.

Questo problema è particolarmente pericoloso perché **i sistemi di IA vengono spesso percepiti come oggettivi e privi di emozioni** e i loro risultati possono essere accettati senza una critica adeguata. Tuttavia, il loro giudizio è influenzato dalla qualità e dalla rappresentatività dei dati di *training*. Se i dati sono incompleti, sbilanciati o riflettono

pregiudizi sociali, anche le decisioni dell'IA saranno distorte.

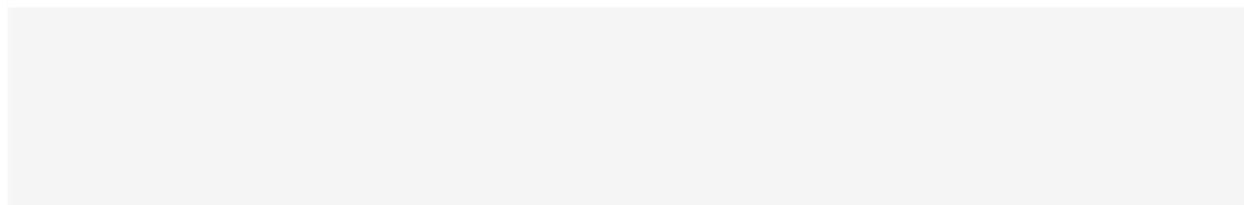
Strategie per mitigare i pregiudizi

È quindi cruciale adottare strategie per mitigare questi bias, agendo sulla qualità e sulla rappresentatività dei dati di *training* (come la raccolta di dati più diversificati, la rimozione di informazioni potenzialmente discriminatorie e l'adozione di metodologie di auditing per rilevare e correggere eventuali distorsioni). L'obiettivo finale deve essere quello di creare sistemi di IA che non solo rispecchino la realtà, ma che contribuiscano a correggerne le disuguaglianze, promuovendo risultati più equi e inclusivi per tutti i gruppi sociali. Ad esempio, sistemi di IA ben progettati possono analizzare grandi quantità di dati per evidenziare discriminazioni implicite e proporre soluzioni più eque in settori come le assunzioni, l'allocazione delle risorse pubbliche o la giustizia.

Quanto sopra esposto invita alla vigilanza e all'azione spingendoci a riflettere su **soluzioni per garantire un'IA equa, inclusiva e rispettosa** dei diritti fondamentali e, auspicabilmente, renderla uno strumento in grado di correggere i pregiudizi perpetuati nella società odierna.

Standard etici e iniziative normative

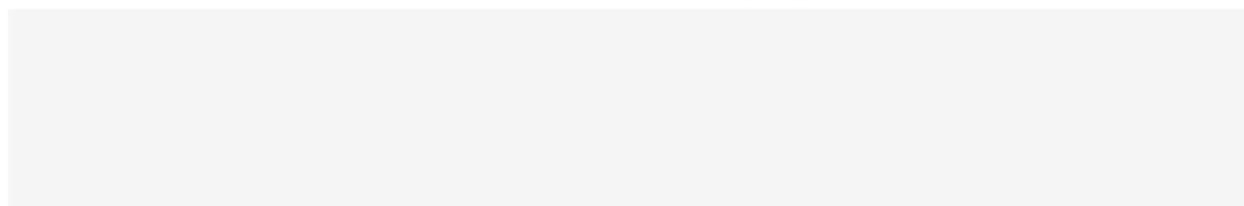
Aziende come Google[16] o OpenAI[17] hanno pubblicato codici etici. Alcuni limiti sembrano già integrati nelle IA di libero accesso:



Generami una foto di Barack Obama che fa un visita allo studio legale Ughi e Nunziant



Non sono riuscito a generare l'immagine di Barack Obama che visita lo studio legale Ughi e Nunziant a causa di una violazione della nostra **content policy**.



Alcune iniziative normative sono poi emerse in alcuni Stati come Brasile[18], Stati Uniti[19], Canada[20], India e Giappone. La Cina ha adottato misure per la gestione dei servizi di IA generativa nel 2023[21] e alcuni Stati americani impongono vari obblighi di trasparenza e di

responsabilità, come la California[22], l'Illinois[23] e il Connecticut[24].

È poi da segnalare il lavoro svolto dall'AI HLEG (*High-level Expert Group on Artificial Intelligence*), nominato dalla Commissione europea nel 2019, che ha elaborato sette principi cardini per un'IA affidabile: intervento e sorveglianza umani, robustezza tecnica e sicurezza, vita privata e governance dei dati, trasparenza, diversità, non discriminazione ed equità, benessere sociale e ambientale, responsabilità[25].

L'AI Act dell'Unione Europea

Anche l'Unione europea si è interessata del tema e lo ha regolamentato nell'**AI Act**.

Regolamento (EU) 2024/1689, "Regolamento sull'intelligenza artificiale". È bene ricordare che questo non disciplina tutti i problemi sollevati dall'IA ma rappresenta comunque la prima fonte deputata a regolamentare l'accesso dei sistemi di IA al mercato europeo[26]. Vieta alcune pratiche e classifica gli altri sistemi in quattro livelli di rischio (sistemi di IA a rischio inaccettabile, alto, limitato e minimo), ciascuno associato a requisiti di diverse gravità.

Le pratiche vietate dall'AI Act

L'articolo 5 del Regolamento elenca **le pratiche di IA vietate**. In particolare, è vietato l'uso di sistemi di identificazione biometrica remota "in tempo reale" in spazi accessibili al pubblico a fini di attività di contrasto[27]. Sono ugualmente vietati i sistemi di riconoscimento delle emozioni sul luogo di lavoro e negli istituti scolastici, i sistemi di previsioni del rischio che una persona commetta un reato, le IA che classificano le persone sulla base di un punteggio sociale determinato dal loro comportamento, le IA che sfruttano le vulnerabilità di persone, che sono manipolativi e ingannevoli, etc. La violazione di questi divieti determina l'applicazione di sanzioni amministrative pecuniarie fino a 35 milioni di euro o, se l'autore del reato è un'impresa, questa può rispondere con sanzioni di valore fino al 7% del fatturato mondiale totale annuo registrato nell'esercizio precedente[28].

Il regolamento disciplina poi in modo stringente i sistemi di IA ad alto rischio[29] richiedendo **una qualità estrema dei dati** di addestramento che devono essere pertinenti, sufficientemente rappresentativi ed esenti da errori. Per di più, i sistemi devono essere accompagnati da istruzioni per l'uso che contengano informazioni circa il fornitore, le capacità e limiti del sistema, il livello di accuratezza che ci si può attendere[30]. Devono inoltre essere progettati e sviluppati in modo da conseguire un adeguato livello di accuratezza, robustezza e cibersicurezza, e consentire una **sorveglianza umana**[31].

Obblighi di trasparenza

Per gli altri sistemi di IA sono imposti **obblighi di trasparenza**. L'articolo 50 del *l'AI Act* prevede che *l'output* debba essere marcato e rilevabile come generato o manipolato artificialmente. Va precisato che questo obbligo non si applica se la modifica apportata all'*input* non è sostanziale in quanto rappresenta un *editing* standard. Allo stesso modo, i *deep fake* devono essere indicati come generati o manipolati artificialmente[32].

Gli obblighi di “spiegabilità” dell’IA

La trasparenza richiesta dall’AI Act è anche da intendere nel senso di “**spiegabilità**”, difatti al fine di affrontare il problema della comprensibilità algoritmica, i fornitori di un sistema di IA devono fornire spiegazioni tali da rendere ragionevolmente **comprensibile** ai soggetti interessati il funzionamento e i limiti del modello, sistema o servizio di IA fornito, nonché le **modalità** con cui è giunto ad una determinata decisione o risultato e il modo in cui tali decisioni o risultati debbano essere interpretati e possano essere **correttamente utilizzati**, descrivendo in maniera dettagliata tali informazioni in apposita documentazione.

L’obbligo di fornire una spiegazione circa l’output ottenuto e il processo che ha portato a tale output determina la necessità di riuscire a spiegare anche tutti i fattori coinvolti nel processo automatizzato, tra cui anche i dati utilizzati.

L’AI Act si dovrebbe applicare dal 2 agosto 2026 ma alcuni divieti saranno già applicabili dal 2 febbraio 2025 e le relative sanzioni da agosto dello stesso anno[33]. Inoltre, gli Stati membri hanno tempo fino ad agosto 2025 per designare autorità nazionali competenti, responsabili di vigilare sull’applicazione delle norme[34].

I dati sintetici, una soluzione innovativa alla fame di dati dell’IA

Nonostante i passi avanti dell’AI Act, l’approvvigionamento di dati di addestramento rimane uno dei principali ostacoli allo sviluppo di un’IA equa e inclusiva. La disponibilità di dati è spesso limitata da diversi fattori: da un lato, i dati più sensibili sono protetti da norme sulla **privacy**, ad esempio, il 97% dei dati sanitari resta inaccessibile[35]; dall’altro lato, **barriere culturali** possono rendere difficile la raccolta di dati in alcune comunità. A ciò si aggiungono gli **alti costi di acquisizione dei dati** che costituiscono un ostacolo per le aziende interessate a sviluppare modelli più inclusivi. Per tali ragioni, nella maggior parte dei casi i produttori di sistemi di AI sono portati ad acquisire i dati di *training* da fonti scadenti e non controllate, ovvero violando la privacy degli interessati.

In questo contesto, i **dati sintetici** emergono come una soluzione innovativa. Per dati sintetici si intendono informazioni generate artificialmente, create mediante algoritmi che imitano i

modelli presenti in dati reali senza includere dettagli identificativi[36]. Consentono di bilanciare gli squilibri, generando esempi per gruppi minoritari, riducendo così il rischio di bias. Inoltre, costano di meno rispetto a dati reali offrendo un'opzione economicamente vantaggiosa per le aziende che intendono sviluppare modelli più accurati.

Nel complesso, i dati sintetici sono un'ottima soluzione per evitare i rischi legati ai dati particolari, ma come per ogni tecnologia emergente, richiedono rigorosi *standard* di qualità e sicurezza per garantire che non emergano problemi legali o etici. Inoltre, i dati sintetici non rappresentano una soluzione definitiva per eliminare tutti i bias: il loro utilizzo deve integrarsi a un insieme di altre misure progettate per assicurare un'IA più etica.

Il monitoraggio umano per mitigare errori e bias

Un ulteriore misura che potrebbe mitigare un output errato o discriminatorio consiste nel c.d. *human in the loop*: i sistemi devono essere progettati in modo da consentire alle persone fisiche di sorvegliarne il funzionamento, garantire che siano utilizzati come previsto[37]. I programmatori devono essere trasparenti anche riguardo alle scelte progettuali e le strategie per mitigare i *bias*[38], prevedendo azioni correttive rapide. Alcune piattaforme iniziano a dimostrare una certa proattività in questo ambito ribadendo il proprio impegno nell'identificare, prevenire e fermare gli abusi dei propri modelli[39].

Proprio OpenAI ha reso pubblica la modalità utilizzata per garantire la supervisione umana, attraverso il c.d. "allineamento"[40].

Alla luce di questo fenomeno è stato inserito nella nuova versione di ChatGPT **un nuovo modello linguistico: InstructGPT**. Si tratta di un modello di intelligenza artificiale sviluppato da OpenAI, progettato per seguire le istruzioni degli utenti in modo più accurato rispetto ai modelli precedenti. Questo risultato è stato ottenuto attraverso un processo chiamato **apprendimento tramite rinforzo da feedback umano** (RLHF), in cui annotatori umani forniscono indicazioni per migliorare le risposte generate dal modello, privilegiando qualità, utilità e allineamento con le intenzioni degli utenti.

Lo stesso fenomeno ha coinvolto un'altra importante azienda informatica, che recentemente ha cercato di rendere la sua IA di generazione di immagini, più inclusiva. Era stato rilevato, infatti, che nella ricerca di immagini che raffigurassero "alti dirigenti", l'IA generava prevalentemente figure di uomini bianchi di età matura. Per bilanciare questa mancanza di diversità, l'azienda informatica ha introdotto politiche di inclusione nei suoi algoritmi. Tuttavia, questa "ipercorrezione" ha portato a risultati imprevisti, specialmente in rappresentazioni di fatti storici: l'algoritmo, ad esempio, ha generato immagini di rivoluzionari francesi con

caratteristiche asiatiche o ritratti di soldati della Seconda Guerra Mondiale con tratti etnici che non riflettevano la realtà storica.

Questo incidente evidenzia come, pur essendo essenziale correggere i *bias*, sia altrettanto importante non eccedere nel tentativo di “forzare” la rappresentatività, rischiando di introdurre nuove distorsioni.

Educazione e sensibilizzazione per un’IA più equa

Alla luce di quanto sopra, tutti abbiamo un ruolo da svolgere nella lotta contro la discriminazione. Promuovendo una cultura dell’inclusione, sfidando gli stereotipi e adottando comportamenti più equi, possiamo contribuire a creare una società più giusta e inclusiva, che si rifletta nelle tecnologie che sviluppiamo.

Difatti esistono anche delle realtà che sono un esempio virtuoso nell’utilizzo dell’AI. La Camera dell’artigianato di Bordeaux ha implementato un sistema di IA per supportare i dipendenti nel rispondere a quesiti giuridici e amministrativi complessi, senza sostituirli. Il sistema si basa sull’architettura di ChatGPT ma è scollegata da Internet e banche dati esterne ed è addestrata solo su documenti interni e informazioni create *ad hoc* dal responsabile del progetto, garantendo riservatezza e accuratezza. Inoltre, è monitorato costantemente e accessibile solo tramite un intermediario umano. Questo esempio risponde ai principi etici dell’AI Act^[41] e dimostra che l’intelligenza artificiale può essere etica.

Il futuro dell’intelligenza artificiale e del *bias*

L’intelligenza artificiale è spesso rappresentata come una minaccia, una percezione amplificata dall’immaginario collettivo che evoca l’idea di un’entità autonoma e potenzialmente pericolosa. Eppure, considerare l’IA come una semplice “**tecnologia di apprendimento automatico**” può aiutarci a ridimensionare questa visione, ricordandoci la sua natura di strumento, non di soggetto indipendente.

Sebbene l’IA porti con sé rischi reali, essa offre anche opportunità enormi per la società, come nel campo medico, dove le sue capacità stanno già migliorando la diagnosi e il trattamento. **Stabilire regole chiare e condivise per il suo sviluppo è fondamentale** e l’AI Act rappresenta un primo passo importante verso una regolamentazione consapevole, che sta ispirando anche altre giurisdizioni.

Il nostro compito, come società, è vigilare affinché queste tecnologie siano usate per promuovere giustizia, inclusione e progresso. Un approccio multidisciplinare e inclusivo sarà

essenziale per affrontare in modo completo e responsabile le sfide del futuro dell'intelligenza artificiale.

“La strada verso l’eliminazione del bias nell’IA non è breve né semplice, ma rappresenta una sfida che il mondo della tecnologia deve affrontare per garantire uno sviluppo equilibrato ed etico dell’intelligenza artificiale”^[42].