L’algoritmo discriminante

Università Roma Tre

10 dicembre 2024

Anno 2054: nella città di Washington non si verifica un omicidio da sei anni.

Il merito è del sistema Precrimine, basato sulle premonizioni dei Precog, tre individui dotati di poteri di precognizione che, prevedendo i delitti, permettono alla polizia di intervenire prima che questi avvengano.

Si esclude la possibilità che i Precog possano sbagliare.

Calato nel mondo algoritmico il senso è che raffinatissimi **strumenti statistici di *pattern-recognition*** se dotati di un numero sufficiente di dati, sono in grado di portare a termine anche i compiti più ardui, in maniera apparentemente infallibile.

L'idea di base è che se c'è un difetto non può che essere umano.

Ma è proprio così? Il difetto può essere soltanto umano e non nel sistema in sé, che si tratti degli algoritmi di *deep learning* o dei Precog?

State v. Loomis (2016)

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) è un software algoritmico utilizzato negli Stati Uniti per valutare il rischio di recidiva.

Sviluppato dalla società **Northpointe, Inc. (ora Equivant)**, COMPAS è impiegato da giudici e sistemi giudiziari per supportare decisioni come concessione della libertà condizionale, cauzione o sentenze. Tuttavia, il software è stato oggetto di numerose critiche per **pregiudizi discriminatori**, specialmente contro persone di colore.

State v. Loomis (2016), presso la Corte Suprema del Wisconsin.

- **Fatti**: Loomis contestò l'uso del punteggio COMPAS nella sua sentenza, sostenendo che il software era opaco e potenzialmente discriminatorio.
- **Decisione**: La Corte confermò l'uso di COMPAS, ma con l'avvertenza che non poteva essere l'unico fattore determinante nella sentenza, sottolineando i limiti dell'algoritmo.

Caso SCHUFA

SCHUFA è una società privata di diritto tedesco che fornisce ai propri partner contrattuali informazioni sulla solvibilità di terzi, in particolare dei consumatori.

A tal fine, stabilisce una prognosi sulla probabilità di un comportamento futuro di una persona ("punteggio"), come il rimborso di un prestito, valutando determinate caratteristiche di tale persona, sulla base di procedure matematiche e statistiche.

La Corte di giustizia europea ha stabilito che questo tipo di pratica costituisce un "processo decisionale individuale automatizzato" ai sensi dell'articolo 22 del GDPR (causa C-634/21, SCHUFA).

SIRI

Nel 2019 l'UNESCO ha pubblicato il documento politico **“I'd blush if I could”** (arrossirei se potessi) che denuncia i pregiudizi di genere nell'assistente vocale di Apple, Siri.

Il titolo era un riferimento alla risposta di Siri quando un utente le avrebbe detto **“Ehi Siri, you’re a bitch!”**

La risposta dell'assistente vocale era **“I'd blush if I could**”.

L'assistente digitale è stato creato nel 2011 e solo dopo la pubblicazione del rapporto dell'UNESCO la risposta è stata corretta.

È stato evidenziato da più parti come i pregiudizi di genere influenzano gli assistenti digitali, in particolare l'uso di assistenti digitali con voce femminile programmati in modo da rafforzare gli stereotipi di genere, ed è stata sottolineata la necessità di includere le donne nello sviluppo di tecnologie guidate dall’intelligenza artificiale per evitare questo tipo di risultati distorti.

Caso Amazon e discriminazioni nei processi di selezione del personale

Un algoritmo di Amazon, utilizzato per valutare i candidati, ha mostrato *bias* di genere.

Il sistema, addestrato su dati storici, **tendeva a sfavorire le candidature femminili per ruoli tecnici**.

Anche se non includeva direttamente il genere come variabile, l'algoritmo sfruttava criteri correlati (proxy discrimination), amplificando disuguaglianze preesistenti.

Caso Deliveroo e algoritmo Frank

Caso Deliveroo e il sistema "Frank" (Tribunale di Bologna, 31 dicembre 2020)

Questa sentenza ha affrontato l'utilizzo dell'algoritmo "Frank" da parte di Deliveroo per la gestione degli slot di lavoro dei rider.

Il tribunale ha stabilito che l'algoritmo era discriminatorio poiché non considerava fattori come malattie o assenze giustificate, penalizzando quindi alcuni lavoratori.

Questa è stata una delle prime decisioni in Europa a dichiarare un algoritmo lesivo del principio di non discriminazione, applicando i principi costituzionali di uguaglianza e tutela dei diritti umani sul lavoro.

Google Ads (Discriminazione razziale e di genere)

Descrizione: L'algoritmo di Google per la pubblicità mirata è stato accusato di mostrare annunci di lavoro per posizioni ben pagate in modo sproporzionato a uomini rispetto alle donne inoltre dava voti negativi a frasi come "sono omosessuale" o "sono ebreo" e voti positivi a "sono un ragazzo eterosessuale".

Critiche:

- Ricercatori hanno scoperto che gli algoritmi di Google tendevano a perpetuare stereotipi di genere e razziali basati sui dati utilizzati per il *training*.

Cos'è un bias?

Il pregiudizio può essere definito come una decisione basata su percezioni ottenute da esperienze culturali e credenze sociali che creano preferenze indesiderate nei confronti di un gruppo, tenendo conto dei suoi attributi unici.

La discriminazione algoritmica si realizza quando nei sistemi di A.I. alcuni errori sistematici e ripetibili distorcono l'elaborazione dei risultati generando output discriminatori.

Il problema può evidenziarsi al momento della progettazione del modello; dell'immissione dei dati e della lettura degli output.

In cosa consiste il rischio di discriminazione algoritmica?

Perchè quando si parla di IA e discriminazione la questione è così allarmante?

Si rischia **di trasporre su modelli automatizzati di larga scala le discriminazioni presenti nelle società.**

La questione è particolarmente complessa perchè da questi *bias* possono generarsi diversi tipi di discriminazione:
Distorsioni del passato;
Proxy discrimination …

In base alla logica *garbage in – garbage out (GIGO)* con cui lavorano gli algoritmi, dati inesatti o non aggiornati non possono produrre altro che risultati inaffidabili.

L’impatto delle decisioni automatizzate sulle libertà fondamentali degli individui: il problema del rischio di discriminazione algoritmica

“I dati con cui vengono addestrati gli algoritmi hanno caratteristiche e problematiche che portano a modelli incorporanti stereotipi e discriminazioni di genere, razza, etnicità e stati di disabilità. **Suprematismo bianco, misoginia e ageismo** etc… sono fin troppo presenti nei dati di addestramento: non solo eccedono la loro presenza in rapporto alla popolazione media ma la creazione di questi modelli, con questi dati, può ulteriormente amplificare i pregiudizi (bias) e danni”.

Questo è il punto di vista di Timnit Gebru, una delle massime autorità in materia di AI, licenziata in tronco da Google dopo la pubblicazione della sua famosa ricerca sui rischi degli algoritmi.

Le difficoltà connesse

Una decisione basata su un pregiudizio non è così facile da identificare nemmeno quando proviene da un cervello umano, figuriamoci da un algoritmo...

Il Consiglio d'Europa (2021) ha evidenziato che i pregiudizi sono difficili da superare perché, dopo il primo giudizio, il cervello umano cerca intenzionalmente di trovare giustificazioni per motivare razionalmente i pregiudizi.

Fonti internazionali

Fonti

Nella settima edizione del rapporto “The AI Index” si evince la complessa portata giuridica dei sistemi di intelligenza artificiale come prioritaria questione politica che in diversi ordinamenti si sta affrontando mediante svariati tentativi di regolamentazione.

Uno sguardo comparatististico evidenzia che gli ordinamenti stanno utilizzando una visione “multistakeholder” aperta al corale coinvolgimento del settore privato, della società civile e degli enti di ricerca, dopo aver riscontrato l’esistenza di un preoccupante livello di divario digitale nel processo di innovazione tecnologica che si sta perfezionando con notevole rapidità, ancorché sulla base di differenziati livelli di implementazione tra i diversi Paesi.

Inoltre si evidenzia formalizza l’esigenza di definire una governance regolatoria dell’IA conforme agli standard internazionali, in grado di prevenire rischi a pregiudizio delle persone.

Fonti - internazionali

- I principi per l'IA dell'OCSE
- La raccomandazione UNESCO sull'etica dell'IA
- La Risoluzione A/78/L49 dell'Assemblea Generale ONU (marzo 2024)
- Convenzione quadro del Consiglio d'Europa sull'intelligenza artificiale, i diritti umani, la democrazia e lo Stato di diritto

Tutte ribadiscono la necessità di preservare i diritti umani, si sancisce un generale divieto di utilizzo dei sistemi di intelligenza artificiale basati sulla codificazione di criteri discriminatori, prevedendosi la generale diffusione di tecnologie emergenti finalizzate, in chiave proattiva, a promuovere l'inclusione.

Qual è la situazione in Europa?

Qual è la situazione in Europa?

L'Unione Europea ha un forte nucleo di valori fondamentali che include i principi di uguaglianza e di non discriminazione.

Essi sono tutelati a più livelli:

Livello primario:

- **Art. 2 TUE** la non discriminazione è tra i valori fondamentali dell'UE. 2 TUE la non discriminazione è tra i valori fondamentali dell'UE;

- **Art. 3 TUE** la promozione della parità tra donne e uomini è considerata uno dei principali obiettivi per lo sviluppo sostenibile del mercato interno;

- **Artt. 20 e 21** della Carta dei diritti fondamentali dell'Unione europea (CFREU) riguardano il principio di uguaglianza davanti alla legge e il principio di non discriminazione.

Qual è la situazione in Europa?

A livello secondario esistono direttive che affrontano la protezione e la promozione dell'uguaglianza e della non discriminazione in diversi settori:

- **Digital Services Act (DSA) e Digital Markets Act (DMA)**: Queste leggi mirano a creare un ambiente online più sicuro, che può indirettamente ridurre i rischi di discriminazione.

Qual è la situazione in Europa?

occupabilità (direttiva 200/78/CE)

servizi sociali (Direttiva 2006/54/CE)

beni e servizi (Direttiva 2004/113/CE)

equilibrio tra lavoro e vita privata per genitori e prestatori di assistenza (Direttiva (UE) 2019/1158)

violenza contro le donne e violenza domestica è la criminalizzazione di alcune forme di cyber-violenza contro le donne (COM(2022)0105);

miglioramento dell'equilibrio di genere fra gli amministratori delle società quotate (Direttiva (UE) 2022/2381)

Qual è la situazione in Europa?

ma...

Mancano disposizioni specifiche contro la discriminazione strutturale, che è proprio quella presente negli ecosistemi di AI, ed è la più difficile da combattere perché si confonde con la neutralità.

Qual è la situazione in Europa?

Regolamento (UE) 1689/2024 (*AI ACT*)

L'UE è recentemente intervenuta sull'intelligenza artificiale adottando il primo regolamento generale in materia.

Si tratta di un testo molto complesso e articolato. I due principali obiettivi perseguiti sono la promozione dell'innovazione e la tutela dei diritti individuali, ma non sempre sono coerenti tra loro.

Come l'AI Act affronta il problema della discriminazione algoritmica

L'AI Act affronta il tema della discriminazione algoritmica attraverso una combinazione di obblighi per la qualità dei dati, la gestione del rischio, la trasparenza e la governance. Questo approccio mira a prevenire danni e discriminazioni sistematiche, responsabilizzando i fornitori di IA e proteggendo i diritti fondamentali.

(Considerando 31, 45, 48, 56, 57, 58, 60, 67, 70, 80…)

1. Approccio orizzontale
2. Approccio basato sul rischio
3. Approccio umanocentrico (l'intelligenza artificiale dovrebbe essere sviluppata considerando il benessere umano come centro di interesse)
4. Sorveglianza umana (art. 14)
5. Obblighi di trasparenza e spiegabilità

In che modo l'AI Act affronta il problema del genere?

- Per i modelli di IA generativa, sono previsti obblighi di trasparenza aggiuntivi, tra cui:
- Valutazione d'impatto.
- **Art. 68, par. 2, lett. c) all'interno del gruppo scientifico di esperti indipendenti deve essere garantita un'equa rappresentanza geografica e di genere.**
- Considerando 27 (che richiama i sette principi etici non vincolanti per l'IA sviluppati dall'HLEG per contribuire a garantire che l'IA sia affidabile ed eticamente valida).
- **Considerando 165 (che fa riferimento alla necessità di promuovere la diversità all'interno dei gruppi di sviluppo, favorendo l'equilibrio di genere).**

In che modo l'AI Act affronta il problema del genere?

L'AI Act rischia di risultare inefficace, per quanto riguarda la protezione dei diritti fondamentali e, in particolare la questione della discriminazione, perché si concentra sulla previsione e sulla prevenzione, ma non sulle misure concrete che possono essere adottate dai cittadini se i loro diritti sono violati da esiti determinati dall'IA.

Tipologie di discriminazioni

Discriminazione diretta: Si verifica quando una persona è trattata in modo meno favorevole rispetto ad altre persone in una situazione simile, a causa di caratteristiche personali come il genere (Art. 2(1)(a) della Direttiva 2006/54/CE).

Discriminazione indiretta: Si ha quando una disposizione, un criterio o una prassi apparentemente neutra mette persone di un determinato genere in una posizione di svantaggio rispetto ad altre (Art. 2(1)(b) della Direttiva 2006/54/CE).

La giurisprudenza della Corte UE

La giurisprudenza della Corte di Giustizia dell'Unione Europea (CGUE) sulle discriminazioni dirette e indirette ha sviluppato principi importanti, ma presenta alcune criticità quando si affrontano questioni legate alle nuove tecnologie.

Inoltre Corte di Giustizia mostra un'eterogeneità nell'interpretazione e nell'applicazione del quadro giuridico europeo in materia di non discriminazione, creando concetti diversi di "uguaglianza" a seconda dei casi, il che porta al problema di definire degli standard comuni di non discriminazione da applicare alle tecnologie guidate dall'IA.

La giurisprudenza della Corte UE - criticità

- La Corte ha adottato un approccio restrittivo nel riconoscere la discriminazione indiretta, richiedendo prove statistiche solide per dimostrare l'effetto discriminatorio di una norma o prassi apparentemente neutrale.
- **La Corte tende a basarsi su modelli tradizionali di discriminazione, concentrandosi sull'intenzionalità del trattamento diverso**. Tuttavia, con le nuove tecnologie, la discriminazione può verificarsi a livello sistemico e non intenzionale (ad esempio, i modelli di IA che penalizzano inconsciamente le donne nei processi di selezione del personale).

Questo è un limite significativo, poiché nel contesto delle nuove tecnologie, le discriminazioni indirette possono essere sottili e difficili da provare.

La giurisprudenza della Corte UE

A causa della mancanza di rinvii pregiuziali, la CGUE non ha ancora affrontato pienamente il problema del “bias algoritmico” che può riprodurre o amplificare disuguaglianze di genere, anche se diversi casi potrebbero far luce su come la CGUE potrebbe decidere in futuro, nel caso in cui sorgesse una controversia sulla discriminazione algoritmica (ad esempio Danfoss; Kelly).

La giurisprudenza della Corte UE

Alcuni casi emblematici mostrano come la Corte affronti la discriminazione indiretta, ma anche come permangano lacune nella sua giurisprudenza in relazione alle nuove tecnologie:

- **Sentenza Danfoss (C-109/88)**: In questo caso, la Corte ha stabilito che, in presenza di un sistema di remunerazione che svantaggia sistematicamente le donne, la discriminazione è presunta e spetta al datore di lavoro dimostrare che esistono criteri oggettivi che giustifichino il trattamento. Tuttavia, nel contesto delle tecnologie avanzate, è più difficile applicare questo principio.

- **Sentenza Chez Razpredelenie Bulgaria (C-83/14):** Qui la Corte ha esteso la nozione di discriminazione indiretta anche a criteri apparentemente neutri che svantaggiano gruppi specifici. Tuttavia, anche in questo caso, la prova della discriminazione deve essere ben fondata, il che può essere complesso nel contesto delle nuove tecnologie, dove il bias può essere nascosto o non facilmente rilevabile.

Conclusioni

Il quadro giuridico dell'UE dispone di validi strumenti a più livelli incentrati sulla tutela dei principi di non discriminazione e uguaglianza.

Inoltre, vi sono chiari sforzi nella *governance* dell'IA per affrontare i problemi legati alle tecnologie guidate dall'IA.

Tuttavia, vi sono delle lacune nelle normative, non ultimo l'AI Act, che rischiano di consentire il perpetuarsi di pregiudizi nei processi algoritmici.

Recenti sviluppi interessanti oltreoceano - New York City Local Law 144

La legge locale 144 della città di New York richiede verifiche annuali indipendenti e imparziali sui pregiudizi degli strumenti automatizzati per le decisioni in materia di lavoro, è conosciuta colloquialmente con nomi quali **NYC Bias Audit Law, NYC Bias Audit Mandate e AEDT Audit Law**.

La legge impone anche requisiti di trasparenza e notifica.

In vigore dal 5 luglio 2023, è la prima legge di questo tipo, ma ha già influenzato proposte simili nello Stato di New York e nel New Jersey.

Cosa serve?

Più una società diventa equa, più questo si rifletterà nei set di dati utilizzati dagli algoritmi.

Cosa serve?

Sensibilizzazione e istruzione (tutela by education)	Promuovere una maggiore diversità nell'industria dell'IA	Quadro giuridico e strumenti per costruire un'IA inclusiva	Audit, standard e certificazioni
• Cambiare la narrazione • Migliorare l'educazione all'IA e alla società, ai pregiudizi e all'etica, per un pubblico tecnico e non tecnico. • Migliorare l'accesso a un'istruzione e a carriere STEM di qualità per le ragazze e le donne.	• Assumere più donne nell'IA Affrontare le disparità in termini di formazione, mantenimento e promozione • Promuovere culture inclusive • Finanziare e investire nelle imprenditrici di IA • Riqualificazione e aggiornamento professionale delle donne	• Aumentare la consapevolezza e formare team sufficientemente eterogenei • Aumentare l'uso di dati intersezionali • Controllo della qualità dei dati • Correzione delle distorsioni dei dati • Pannelli etici • Governance Audit interni sull'equità di genere	• Principi di parità di genere nell'IA • Valutazioni d'impatto algoritmico (AIA) • Audit degli strumenti di auditing • Verifica della manutenzione dello standard IEEE P7003 sui pregiudizi algoritmici • Due diligence

Grazie per l’attenzione!

Cristina Grieco
c.grieco@unimc.it