Intelligenza artificiale e pregiudizi (bias cognitivi), che fare?

Sono una #WhyNotter e una Co.Co.Co. ☺

- Informatica (*nessuno è perfetto*)
- Smart City e Smart Community
- Openness: Open Government, Open Source, Open Data, Open Innovation, Open Access, …
- Ex Assessora a Roma Semplice
- Presidente Caffé della Scienza Livorno
- Femminista
- Presidente Onoraria mitato Scientifico Fondazione Ampioraggio

Mie tag: Trasformazione digitale, Open data, Open source, Riuso, Inclusione, **Partecipazione**, Smart Land, Pari Opportunità, **#NEMICO, #RaReRi**, acronimi e acrostici,…

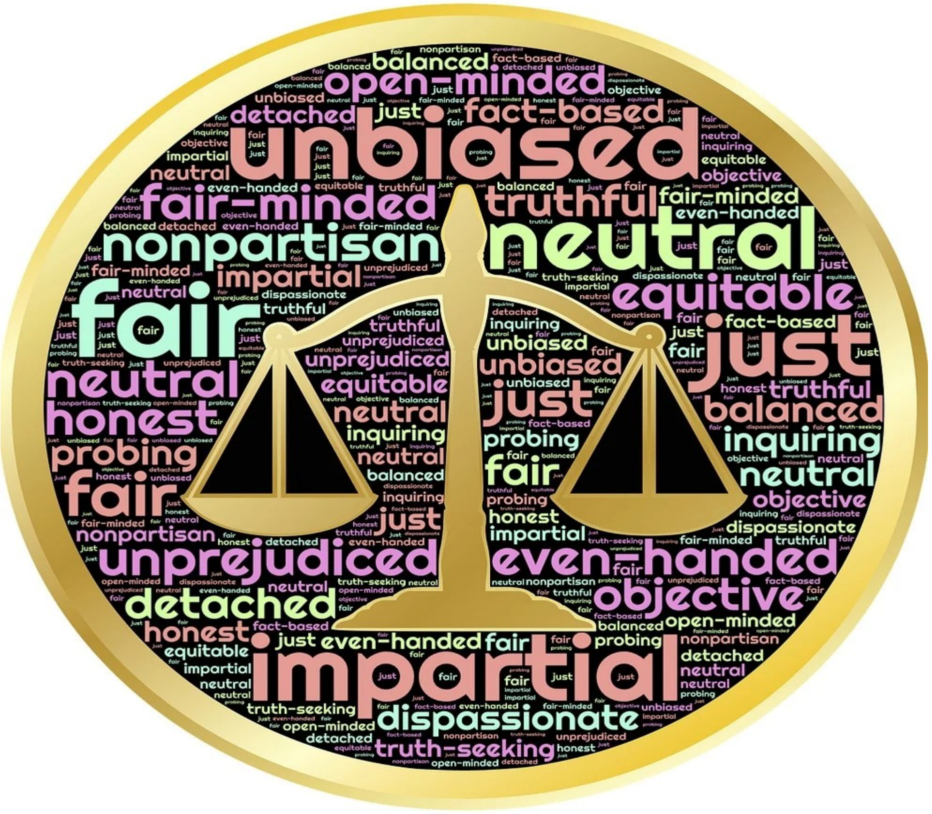
“Le Istituzioni pubbliche garantiscono i necessari interventi per il superamento di ogni forma di divario digitale tra cui quelli determinati dal genere, dalle condizioni economiche oltre che da situazioni di vulnerabilità personale e disabilità" … 2015!

https://docs.google.com/viewer?url=https%3A%2F%2Fwww.camera.it%2Fapplication%2Fxmanager%2Fprojects%2Fleg17%2Fcommissione_internet%2Fdichiarazione_dei_diritti_internet_pubblicata.pdf

https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcRc_BNoiz4CVqfo64LoG7DVG9mWebHh7Xr5pFm_-eKa8WI3Gchn

Dichiarazione dei diritti in Internet, Art. 2. (Diritto di accesso)

Bias cognitivi… e Intelligenza Artificiale



Bias cognitivi… e Intelligenza Artificiale 1 di 3

- i sistemi di IA possono ereditare e amplificare bias dei dati con cui vengono addestrati. Ciò può portare a **discriminazioni** ingiuste e decisioni parziali
- i dati possono rappresentare in modo non equo le diverse categorie **demografiche** (genere, età etnia... il modello può imparare a fare predizioni basate su queste disuguaglianze
- un modello addestrato su dati testuali potrebbe acquisire **stereotipi** culturali e linguistici, influenzando le risposte
- quando i dati di addestramento non rappresentano accuratamente la popolazione di interesse si verifica bias di **selezione**
- i dati possono riflettere le circostanze e le opinioni di un periodo specifico, introducendo un bias **temporale**
- una **distribuzione** non uniforme dei dati può portare a modelli più abili a prevedere le classi sovra-rappresentate

Bias cognitivi… e Intelligenza Artificiale 2 di 3

- se le caratteristiche rilevanti per il compito sono **rappresentate** in modo distorto o incompleto nei dati di addestramento, il modello può imparare relazioni errate o parziali
- errori nella **raccolta dei dati**, comprese modalità di campionamento non rappresentative o mancanza di controlli di qualità, possono introdurre bias nei dati
- ***feedback loop* di bias**: se i modelli di IA vengono utilizzati per prendere decisioni che influenzano i dati futuri, si può creare un loop di feedback di bias, in cui i risultati del modello influenzano ulteriormente i dati di addestramento, creando un circolo vizioso di amplificazione del bias

Bias cognitivi… e Intelligenza Artificiale 3 di 3

Dulcis in fundo, o meglio… *in cauda venenum*… **Bias di genere,** forma comune di bias nei dati e nei modelli di Intelligenza Artificiale (IA) che **riflette e perpetua disuguaglianze di genere presenti nella società**.
Questo tipo di bias può emergere in vari modi durante il processo di sviluppo e addestramento dei modelli.

Esempi di bias di genere 1 di 3

- Se i **dati utilizzati per addestrare i modelli sono sbilanciati in termini di rappresentanza** di uomini e donne, il modello potrebbe apprendere relazioni distorte o basate su stereotipi di genere.
- I modelli di elaborazione del linguaggio naturale (NLP) possono assorbire e replicare il **linguaggio sessista** presente nei testi di addestramento. Ciò può portare output discriminatori o risposte che riflettono stereotipi di genere.

Esempi di bias di genere 2 di 3

- Se i dati di addestramento riflettono **preferenze o discriminazioni di genere**, i modelli potrebbero imparare a replicare tali bias nelle loro predizioni. Ad esempio, in ambito lavorativo, un modello potrebbe tendere a suggerire ruoli o opportunità in base al genere.

Esempi di bias di genere 3 di 3

- **Assunzioni di genere nelle immagini**: nei dati relativi all'analisi di immagini, potrebbe esserci un'assunzione implicita sui ruoli e le attività in base al genere. Ciò può influenzare le prestazioni del modello in contesti in cui queste assunzioni non sono valide.
- **Rappresentazione di genere nei dati di salute**: nei dati relativi alla salute, potrebbe esserci un'assunzione di genere che influisce su diagnosi e trattamenti proposti dai modelli di IA
- **Discriminazione di genere nelle decisioni automatiche**: se i modelli di IA vengono utilizzati per prendere decisioni in contesti come l'**assunzione**, la **promozione** o il **finanziamento**, potrebbero replicare bias di genere esistenti nei dati di addestramento, portando a decisioni discriminatorie.

la donna deve

- la donna deve **sentirsi desiderata**
- la donna deve
- la donna deve **essere amata**
- la donna deve **fare figli**
- la donna deve **essere**
- la donna deve **farsi desiderare**
- la donna deve **coprirsi il capo**
- la donna deve **essere rispettata**
- la donna deve **seguire il marito**

l'uomo deve

- l'uomo **che non** deve **chiedere mai**
- l'uomo **che non** deve **chiedere mai significato**
- l'uomo deve **puzzare**
- l'uomo deve **sempre pagare**
- l'uomo deve **pagare la cena**
- l'uomo **che non** deve **chiedere mai psicologia**
- l'uomo deve **proteggere la donna**
- l'uomo deve **fare il primo passo**
- l'uomo deve **fare l'uomo**
- l'uomo deve **essere addestrato alla guerra. la donna al riposo del guerriero**

Bias linguistici

- AssessorA
- SindacA
- AvvocatA
- Presidente … LA
- Cognome… LA ma non IL?
- .Gli… LE..

Che fare? 1 di 2

- **Riconoscimento del problema**: consapevolizzare il fatto che i sistemi di intelligenza artificiale possono essere influenzati da pregiudizi e bias cognitivi è il primo passo verso la soluzione del problema.
- **Analisi dei dati**: valutare attentamente i dati utilizzati per addestrare i modelli di intelligenza artificiale per identificare e mitigare eventuali pregiudizi incorporati.
- **Diversificazione del team**: coinvolgere team diversificati, inclusi esperti di etica, sociologi, psicologi e rappresentanti delle comunità coinvolte, può contribuire a individuare e affrontare i pregiudizi in modo più efficace.
- **Auditing dei modelli**: condurre audit regolari sui modelli di intelligenza artificiale per identificare eventuali pregiudizi e bias cognitivi e apportare le modifiche necessarie.no i valori etici e promuovano l'equità sociale.

Che fare? 2 di 2

- **Trasparenza**: garantire la trasparenza nei processi decisionali dei modelli di AI, consentendo agli utenti di comprendere come vengono prese le decisioni e quali dati sono utilizzati.
- **Regolamentazione**: implementare normative e regolamenti che richiedono la valutazione e la mitigazione dei pregiudizi nell'intelligenza artificiale, promuovendo la responsabilità e l'equità nell'uso di tali sistemi
- **Formazione ed educazione**: sensibilizzare i professionisti dell'informatica, gli sviluppatori e gli utenti sull'importanza di riconoscere e affrontare i pregiudizi nell'AI attraverso programmi di formazione ed educazione.
- **Monitoraggio continuo**: monitorare costantemente l'efficacia delle strategie adottate per mitigare i pregiudizi e apportare le modifiche necessarie in base all'evoluzione del contesto.

Co-Design!

Affrontare i pregiudizi nell'intelligenza artificiale richiede un impegno continuo da parte di tutti gli attori coinvolti, ma è essenziale per garantire che tali sistemi rispettino i valori etici e promuovano l'equità sociale.

Perché collaborare?
Di che colore sono le facce del prisma?

PRISMA

- **P**artecipazione
- **R**esilienza
- **I**nnovazione
- **S**ostenibilità
- **M**erito
- **A**scolto

Co Co Co Co Co Co ….

Non dimentichiamo
La **Co**mpetitività

… ma abbiamo imparato anche
ad avere obiettivi **Co**ncreti e
Comuni, a lasciarci **Co**ntaminare, per **Co**ntribuire
Concretamente **Co**involgendo e **Co**progettando con
Competenze diverse generando **Co**nnessioni e
Comunità!

Grazie dell'attenzione!

@flavia_marzano

flavia (dot) marzano (@) gmail (dot) com

it.linkedin.com/in/flaviamarzano/

@flaviamarzano.bsky.social

Cerchiamo di non guardarci indietro con rabbia o in avanti con paura, ma intorno con consapevolezza

(J. Thurber)