CATÓLICA
FACULTY
OF LAW
ESCOLA DO PORTO

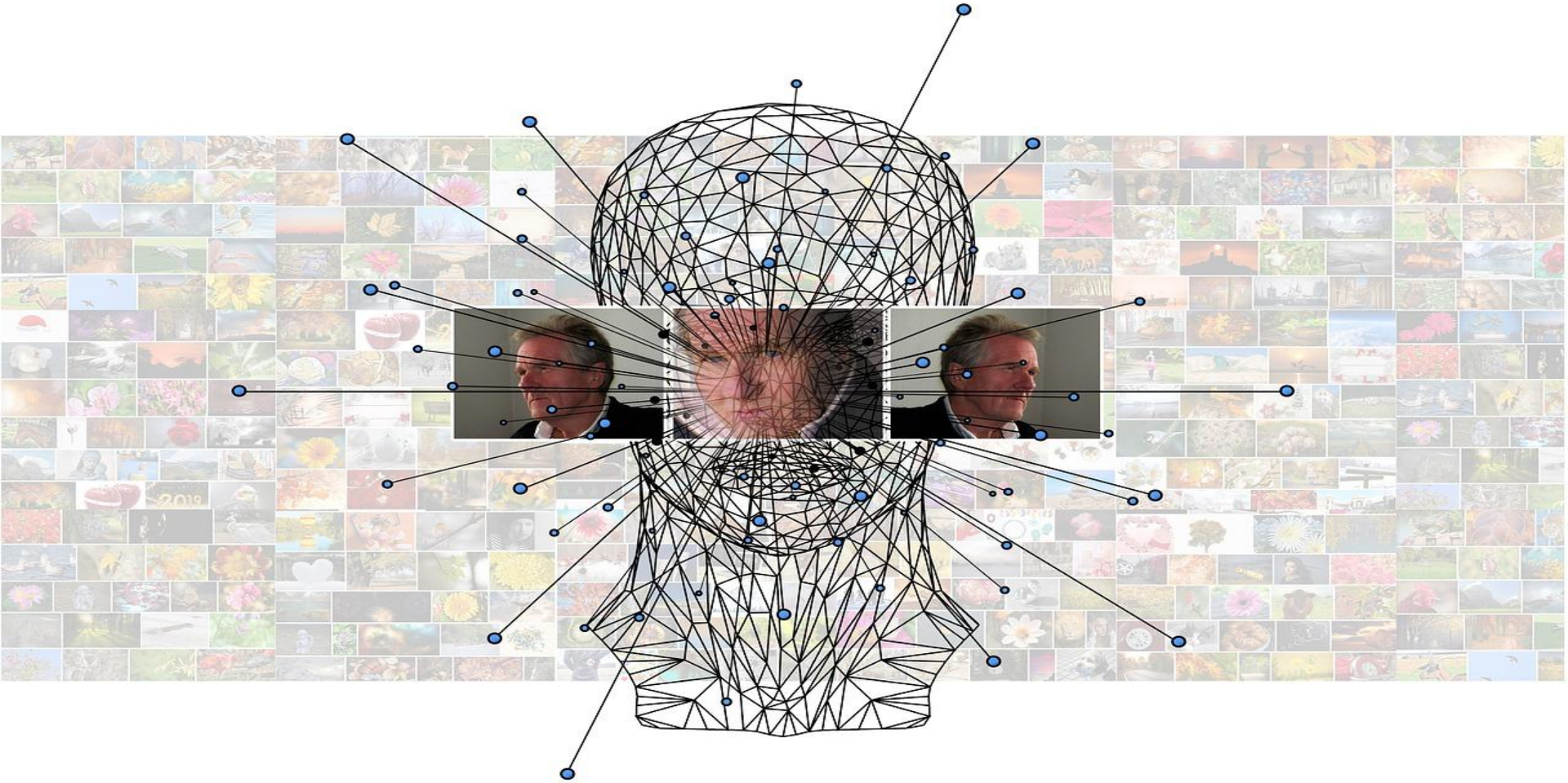
Funded by
the European Union

La fine dell'inganno? - Contrastare la discriminazione algoritmica nell’era digitale

Catarina Santos Botelho
cbotelho@ucp.pt

Cosa sono gli algoritmi?

- Gli algoritmi sono procedure codificate attraverso le quali i dati di input, seguendo una sequenza logica e matematica, si trasformano in un output



- In questa fase, gli algoritmi sono complessi e sfidano la comprensione umana
- Gli algoritmi possono essere classificati in algoritmi basati sulla conoscenza e algoritmi di apprendimento automatico (*Machine-Learning* ML)
- I sistemi **basati sulla conoscenza** sono i tradizionali algoritmi a diagramma di flusso che richiedono un codice manuale
- I **ML** sono algoritmi che alimentano grandi quantità di dati con variabili di output affinché l'algoritmo stesso "impari senza essere esplicitamente programmato"

Perché questo tema è importante?

I processi decisionali automatizzati sono una realtà sempre più diffusa

- Disposizioni in materia di assistenza sanitaria,
- Assegnazione del vaccino,
- Pene detentive,
- Controllo della polizia,
- Selezione dell'audit dell'IRS,
- Conteggio dei voti,
- Determinazioni sui prestiti,
- Domande di lavoro,
- Assegnazione al lavoro,
- Richieste di prestazioni sociali,
- Concessione di visti di immigrazione,
- Selezioni per l'ammissione alla scuola dell’infanzia o all'università
- Prezzi,
- Tutela dei consumatori,
- Marketing di nicchia.

Quali sono le domande importanti?

- Cosa succederà alla democrazia e alla tutela dei diritti umani nella nuova "architettura socio-tecnica di Internet"?
- Gli algoritmi possono intensificare i contraccolpi democratici?
- L'intervento umano nei processi decisionali è ancora fondamentale per la tutela dei diritti umani o è una precauzione inutile?

Discriminazione diretta / Discriminazione indiretta

- **Diretta**: quando una persona è trattata meno favorevolmente di un'altra a causa di determinate caratteristiche in un determinato settore protetto

- **Indiretta**: quando una disposizione, un criterio o una prassi, apparentemente neutri, mettono in difficoltà una persona appartenente a un gruppo protetto senza una giustificazione accettabile

Trattamento differenziato / Impatto differenziato

- Titolo VII della Legge sui diritti civili del 1964 (USA)

- *Trattamento differenziato*: quando il datore di lavoro tratta il dipendente in modo diverso a causa di una caratteristica protetta.
- *Impatto differenziato*: una pratica che a prima vista sembra neutra e non discriminatoria, ma che in realtà arreca svantaggio in modo sproporzionato ad un gruppo protetto.

Criteri per i gruppi protetti

- I gruppi protetti sono fissi o si possono ad essi aggiungere altri "assi di identità socialmente salienti"?
- Possono lo stato socio-economico o il livello di istruzione essere presi in considerazione ai fini delle proposte antidiscriminatorie? E che dire dell'obesità, delle malattie croniche o del tipo di browser?
- Le proposte legislative in materia di antidiscriminazione, in particolare le liste dei motivi e/o delle eccezioni, dovrebbero essere aperte, chiuse o ibride?

Pregiudizi sugli algoritmi di ML:

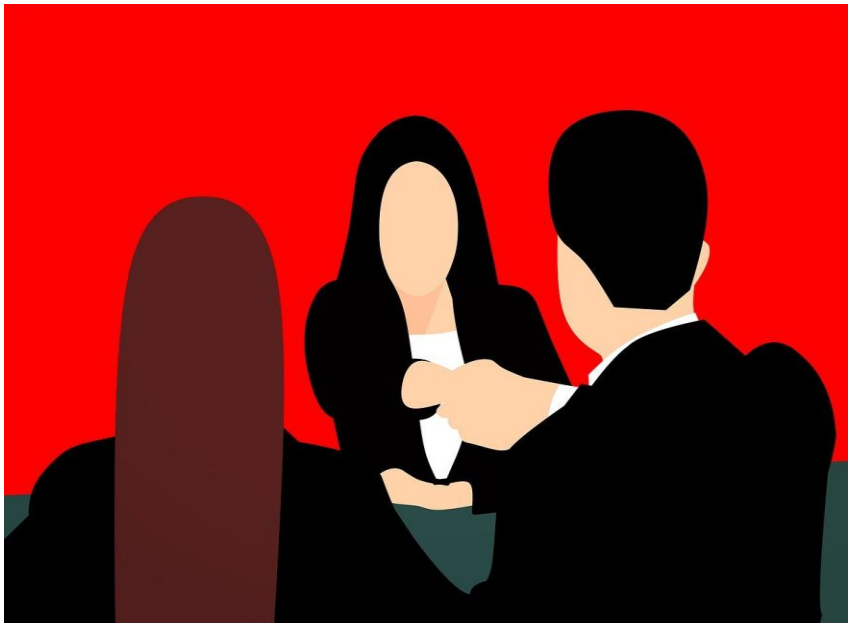
- *Pregiudizi conseguenti*, quando i creatori dell'algoritmo sono intenzionalmente o involontariamente prevenuti, e tali pregiudizi si riprodurranno quindi nel meccanismo di elaborazione dei dati

- *Dati campione distorti*, in cui i dati campione o di addestramento forniti sono distorti

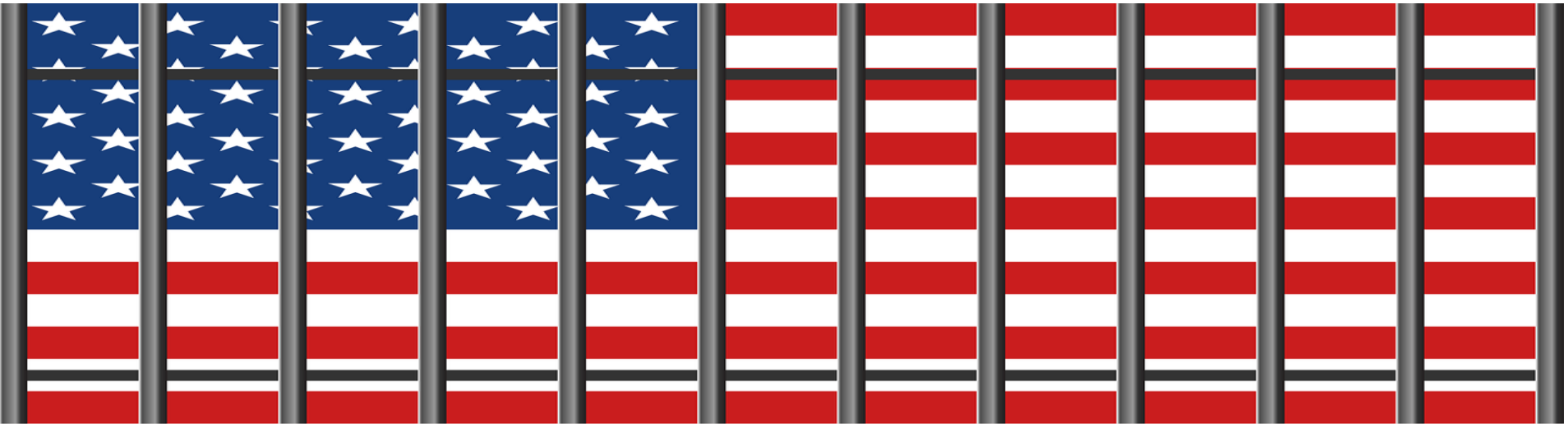
- *Pregiudizi di risultato* o *pregiudizi del circolo vizioso*, che si verifica anche quando l'algoritmo viene addestrato con dati rappresentativi, a causa delle disuguaglianze sociali incorporate nei dati

Esempi di impatto disparato algoritmico e discriminazione statistica

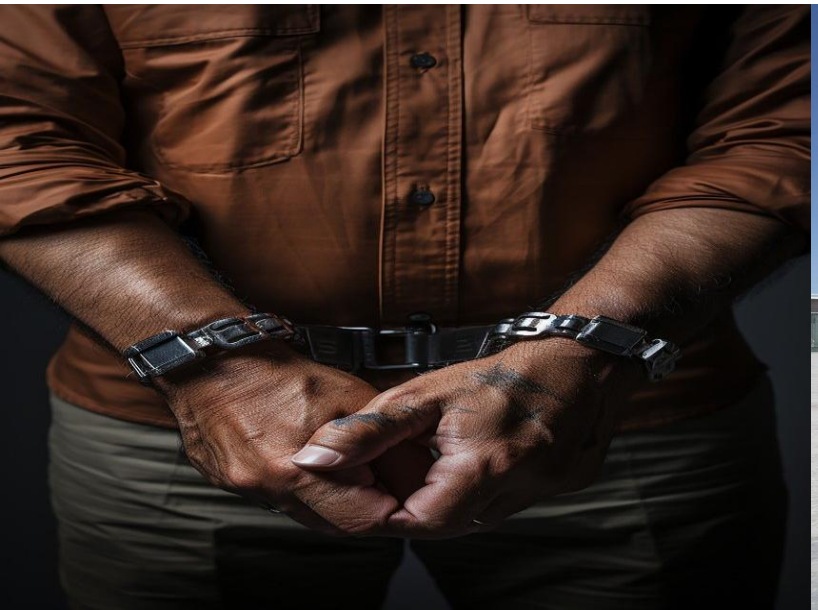
- ***Caso Amazon***:
- Amazon ha sviluppato un sistema di ML per identificare i *futuri dipendenti sviluppatori di software*. L'algoritmo classificava i candidati in base alla produttività prevista. Poiché l'algoritmo era stato addestrato utilizzando i dati di assunzione esistenti, che riflettevano l'attuale predominio maschile nell'industria tecnologica, i candidati con indicatori femminili nei loro curriculum (ad esempio, club di tennis femminile) sono stati discriminati indipendentemente dalle loro qualifiche per il lavoro.



- **Caso COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*):**
- Utilizzando i dati di una contea statunitense, una società ha sviluppato uno strumento di valutazione del rischio noto come COMPAS per prevedere la recidiva di detenuti e imputati
- Alla domanda "questa persona sarà recidiva se le venisse concessa la libertà vigilata?" l'algoritmo ha confrontato le caratteristiche della persona con quelle di coloro che hanno commesso reati durante la libertà vigilata



- Il problema segnalato da ProPublica nel 2016 era che COMPAS utilizzava la sua previsione di riarresto come un indicatore di recidiva
- Il COMPAS è stato accusato di produrre risultati distorti dal punto di vista razziale, dal momento che i falsi positivi (che prevedono un alto rischio di recidiva per l'autore del reato non recidivo) per gli imputati afroamericani risultavano il doppio di quelli degli imputati di origine caucasica



L'illusione dell'oggettività algoritmica

- "Avversione algoritmica" / "AI come mostro"

contro

- Superpoteri erculei / campioni di obiettività

Discriminazione algoritmica endogena

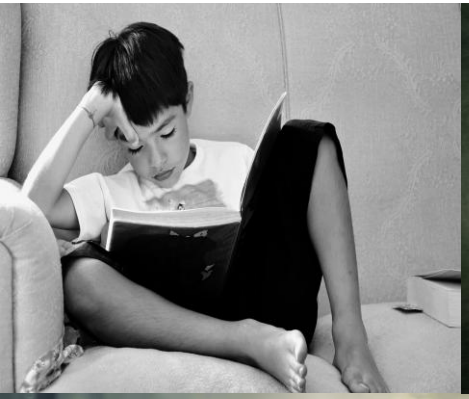
- La discriminazione sta nella programmazione stessa dell'algoritmo
- Più trasparenza: impedire la migrazione del processo decisionale umano discriminatorio a quello degli algoritmi
- Può la trasparenza essere considerate la nuova trappola del consenso?

Discriminazione algoritmica esogena

- Il problema non è la "scatola nera" in sé, ma il *mondo* reale in cui operano gli algoritmi
- La discriminazione è il risultato di pregiudizi impliciti, distorsioni statistiche o travisamenti di gruppi storici che si sono fatti strada attraverso un ML presumibilmente asettico
- L'idea è quella di "andare oltre la semplice *non discriminazione* per passare ad algoritmi attivamente *antidiscriminatori*" (Bornstein)

Uguaglianza sostanziale

- ***L'azione affermativa algoritmica*** può essere ottenuta attraverso diverse misure come l'incorporazione di metriche di equità nei processi di selezione dei modelli, o attraverso l'equità di gruppo utilizzando approcci di "parità demografica" o "parità statistica"






Algoritmi di regolazione e debiasing

- Poiché gli algoritmi sono incomprensibili per gli esseri umani è necessario un ulteriore sforzo per scoprire i risultati discriminatori
- Date le circostanze, i politici, i tribunali, le agenzie, le autorità pubbliche e i datori di lavoro non dovrebbero adottare una posizione acritica e di eccessiva fiducia negli algoritmi
- Si raccomanda vivamente, quindi, una vigilanza attiva

1. Controllo algoritmico

- L'Atto europeo sull'IA mira in modo pionieristico ad una intelligenza artificiale etica e incentrata sull'uomo, dove i risultati dei sistemi di IA siano tracciabili, spiegabili e trasparenti

- I sistemi di intelligenza artificiale sono classificati in base al loro livello di rischio potenziale (articolo 5):

- Vietato a priori
- Ad alto rischio
- A basso rischio

- Test di proporzionalità

- Per ridurre al minimo l'impatto potenziale dei sistemi di IA ad alto rischio, i responsabili dell'installazione devono condurre **valutazioni d'impatto sui diritti fondamentali** (*Fundamental Rights Impact Assessment - FRIA*) prima di installare un sistema di IA ad alto rischio (articolo 27)

Articolo 6 (Legge europea sull'AI)

3. In deroga al paragrafo 2, un sistema di IA non è considerato ad alto rischio se non presenta un rischio significativo di danno per la salute, la sicurezza o i diritti fondamentali delle persone fisiche, anche nel senso di non influenzare materialmente il risultato del processo decisionale.

Il primo comma si applica quando è soddisfatta almeno una qualsiasi delle condizioni seguenti:

a) il sistema di IA è destinato a eseguire un compito procedurale limitato;

b) il sistema di IA è destinato a migliorare il risultato di un'attività umana precedentemente completata;

c) il sistema di IA è destinato a rilevare schemi decisionali o deviazioni da schemi decisionali precedenti e non è finalizzato a sostituire o influenzare la valutazione umana precedentemente completata senza un'adeguata revisione umana;

d) il sistema di IA è destinato a eseguire un compito preparatorio per una valutazione pertinente ai fini dei casi d'uso elencati nell'allegato III.

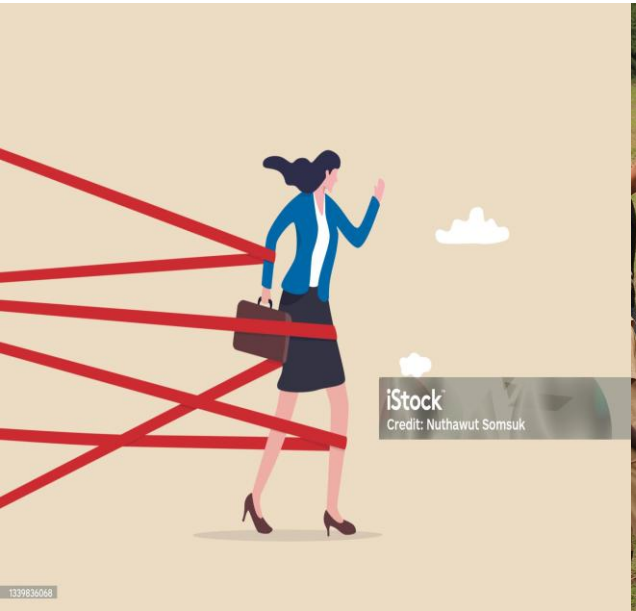
2. Blocco delle informazioni: "equità per cecità?"

- Il blocco delle informazioni personali sulle categorie protette nei dati di formazione previene i risultati discriminatori?
- Per ottenere risultati senza distinzione di razza, disabilità, genere, reddito o codice postale, potrebbe non essere sufficiente (e nemmeno saggio) rimuovere semplicemente gli attributi sensibili dalla considerazione.

- I sistemi di ML eccellono nel "rilevamento di schemi e nell'identificazione e sfruttamento di indicatori non ovvi per variabili sensibili omesse".

- Come possiamo rivelare "strati su strati di specchi e indicatori" e come possiamo sapere quali *indicatori* (sostituti funzionali) contengono informazioni preziose?

- Un ulteriore fattore di complicazione è che la parità di razza e di genere può ancora trascurare *le differenze intersettoriali* all'interno di questi gruppi alla luce di altre dimensioni, come le disabilità, l'età o la nazionalità.



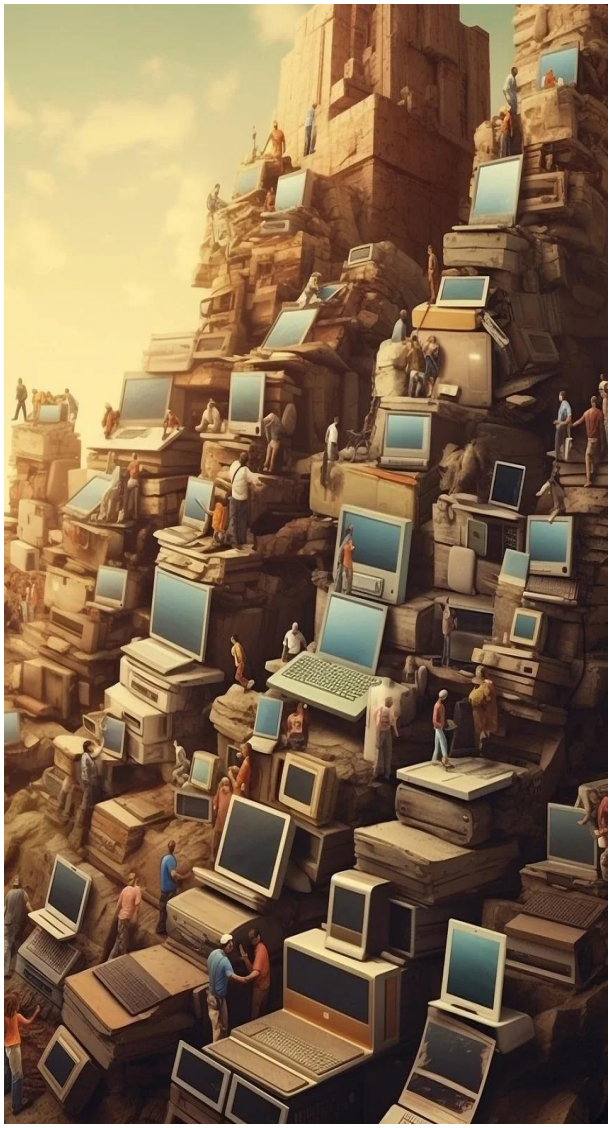
3. Regolare e modellare i dati attraverso la progettazione algoritmica

- Il Regolamento generale sulla protezione dei dati (GDPR) richiede che le attività di trattamento dei dati siano trasparenti (*Consideranda* 39, 58 e 78 del GDPR e articoli 5, paragrafo 1, lettera a), e 12 del GDPR), garantisce il diritto di accesso e il diritto di essere informati dell'esistenza di un processo decisionale automatizzato (articoli 13-15), nonché il diritto di non essere soggetti esclusivamente a un processo decisionale automatizzato (articolo 22), e ha introdotto la valutazione d'impatto sulla protezione dei dati, per valutare i rischi per i diritti e le libertà degli interessati e i rischi dell'attività di trattamento (articolo 35). Tuttavia, questo diritto non si applica quando l'individuo ha fornito un consenso esplicito per la decisione automatizzata (articolo 22, paragrafo 2, lettera b) del GDPR).

A questo punto, più che regolare gli algoritmi, è importante regolare i dati che vengono forniti agli algoritmi.

L'obiettivo è raccogliere "dati *equi*" che possano controbilanciare i pregiudizi della società.

Invece di bloccare i dati, alcuni studiosi propongono *di modificarli*. La pre-elaborazione dei dati può essere ottenuta regolando i dati di input in generale o regolando la variabile obiettivo (ad esempio, il genere).



4. Strategie ex post per riparare la discriminazione: diritti individuali o diritti di gruppo?

- Tribunali (attraverso il controllo giurisdizionale)
- Le autorità preposte alla protezione dei dati (ad esempio, attraverso le valutazioni d'impatto sulla protezione dei dati)
- "FDA per gli algoritmi"?
- “Un diritto individuale di contestare le decisioni dell'IA"?

I punti deboli degli algoritmi e l'importanza della centralità dell'uomo

- Esiste il "diritto a una decisione umana"?
- L'IA può imitare e persino superare l'intelletto umano. Tuttavia, il ragionamento induttivo, la cognizione umana e la presenza umana nell'esercizio della *discrezione* (che è diversa dall'arbitrarietà) o del giudizio per il caso specifico sono ancora necessari.

Perché?

- Perché solo la persona umana può comprendere appieno la realtà. L'IA esegue operazioni logiche che sono astrazioni della realtà (il mondo esterno o il mondo reale). Così facendo, l'IA crea la sua realtà all'interno della "realtà", anche se si tratta sempre di una realtà ristretta che non può raggiungere ontologicamente la sua *essenza*. Al contrario, gli esseri umani *sperimentano e abbracciano la realtà*.

- Non si tratta di una futile competizione tra uomini e macchine. Non abbiamo bisogno dell'intervento umano nell'IA perché "noi umani" siamo migliori.
- Abbiamo bisogno di una guida umana perché non possiamo escludere l'umanità, l'etica e la morale, dai processi decisionali pubblici e privati che possono incidere profondamente sulle nostre vite.