Bias algoritmici: pregiudizi discriminatori nell'addestramento dei sistemi di intelligenza artificiale

UNIVERSITÀ "MAGNA GRAECIA"

23 novembre 2024

A cura dell'Avv. Giulia Bifano

Senior Associate

Baker McKenzie.

Agenda

1. Cos’è il bias nell’AI
2. Esempi di bias nei sistemi di HR
3. Meccanismi per identificare e correggere i bias nell’AI
4. Esempi di best practices
5. Implicazioni etiche e legali nell’utilizzo di AI nei luoghi di lavoro
6. Necessità dell’intervento umano
7. Q&A

1 Cos’è il bias nell’AI

Il bias nell'intelligenza artificiale (AI) si riferisce a distorsioni o pregiudizi presenti nei sistemi di AI, che possono influenzare negativamente le decisioni e i risultati prodotti. Questi bias spesso derivano da dati di addestramento contenenti pregiudizi storici o da scelte progettuali degli algoritmi stessi.

Bias di selezione

Bias di conferma

Bias di rappresentazione

Bias algoritmico

2 Esempi di bias nei sistemi di HR

Gli algoritmi utilizzati nei sistemi di gestione delle risorse umane (HR) hanno mostrato casi di bias che hanno portato a discriminazioni evidenti, sottolineando la necessità di adottare misure correttive e preventive.
Gli algoritmi di selezione utilizzati da aziende di alto profilo hanno evidenziato pregiudizi contro determinate categorie di candidati.
Ecco alcuni casi concreti:

Amazon e il bias di genere

- Il sistema di reclutamento AI di Amazon, addestrato su dati storici, penalizzava sistematicamente le candidature femminili per ruoli tecnici. Il motivo era che i dati riflettevano una preferenza storica per candidati uomini.

COMPAS e il bias razziale

- Anche se non specifico agli HR, il sistema COMPAS, utilizzato per valutare la recidività, ha mostrato un forte bias contro gli afroamericani, spesso sovrastimando il rischio rispetto ai detenuti bianchi. Sebbene progettato per essere neutrale, il modello riproduceva pregiudizi presenti nei dati.

Software di analisi del linguaggio

- Sistemi che analizzano il linguaggio nei CV o durante colloqui spesso mostrano bias culturali. Ad esempio, algoritmi possono favorire candidati che utilizzano termini "mainstream" e penalizzare quelli con terminologie diverse, spesso legate a minoranze linguistiche o culturali.

Conseguenze per le aziende e i dipendenti

Gli effetti dei bias negli HR non sono solo dannosi per i candidati, ma possono avere implicazioni gravi anche per le aziende.

- **Danno all'immagine aziendale**:

Scandali legati a bias algoritmici possono danneggiare seriamente la reputazione dell'azienda. Ad esempio, il caso Amazon ha sollevato un dibattito globale sulla fiducia negli algoritmi e sulla responsabilità delle aziende.

- **Perdita di talenti:**

Se un sistema penalizza candidati qualificati (ad esempio, donne o minoranze), l'azienda rischia di perdere talenti preziosi, compromettendo la diversità e l'innovazione all'interno dell'organizzazione.

- **Rischi legali:**

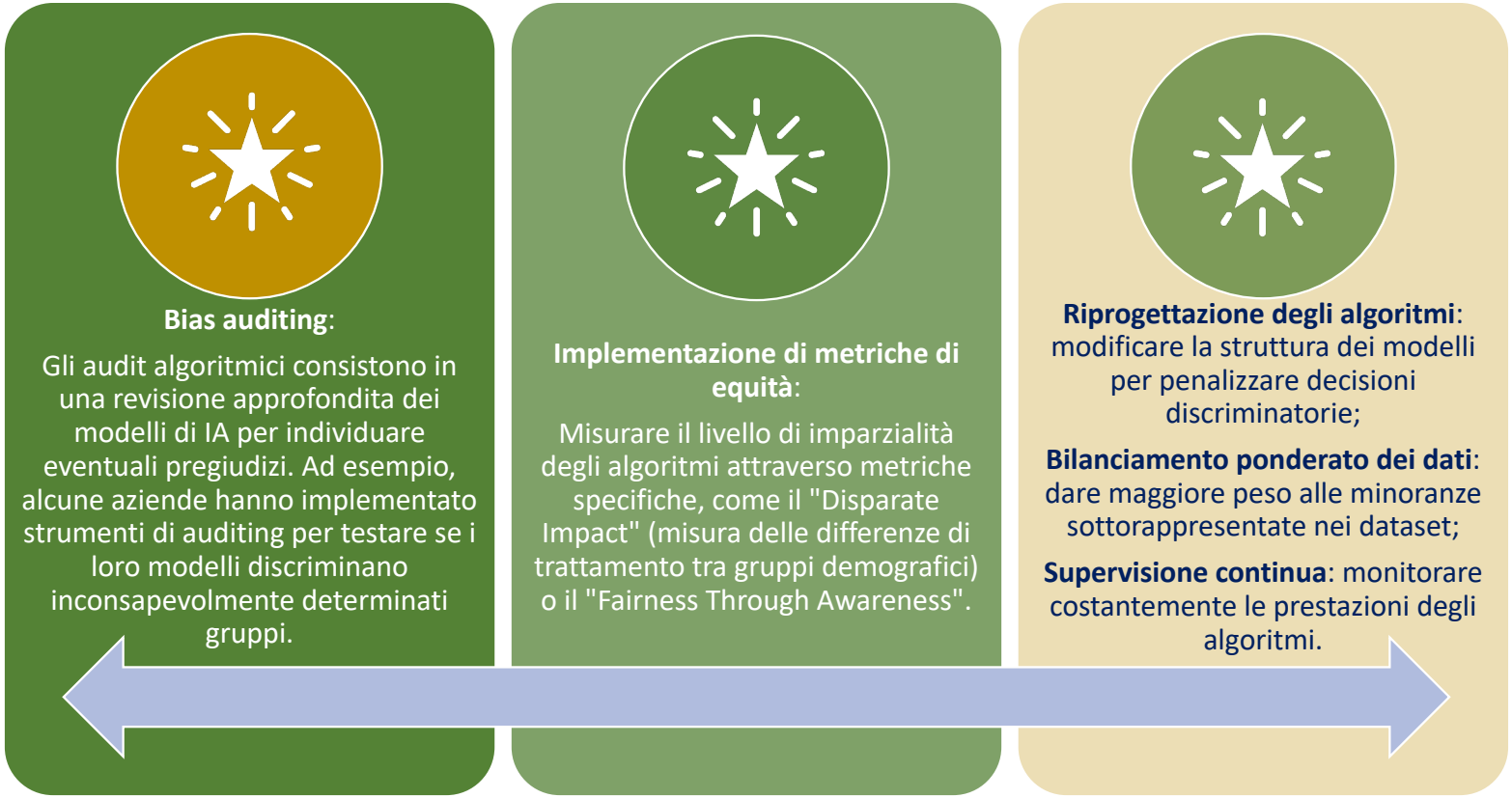
Discriminazioni sistematiche possono portare a cause legali, multe e sanzioni, in particolare in paesi con normative stringenti contro la discriminazione (ad esempio, negli Stati Uniti o in Europa con il GDPR).

3 Meccanismi per identificare e correggere i bias nell'AI

Per mitigare i rischi legati ai bias nell'intelligenza artificiale, è fondamentale implementare strategie che identificano, monitorano e correggono pregiudizi nei modelli di IA. Questi approcci si concentrano su dati inclusivi, processi trasparenti e interventi correttivi.
Uno dei principali fattori che causano bias negli algoritmi è l'uso di dati di addestramento non rappresentativi. Per prevenire queste distorsioni, le aziende devono:

- **Ampliare i dataset utilizzati**:
  - Garantire che i dati raccolti includano una rappresentazione equilibrata di genere, etnia, età e altre caratteristiche demografiche.
  - Raccogliere dati da fonti diversificate per ridurre l'influenza di pregiudizi culturali o geografici.
- **Valutare la qualità dei dati esistenti**:
  - Analizzare i dataset per individuare pregiudizi preesistenti e correggerli prima di utilizzarli per l'addestramento dei modelli.
- **Aggiornare regolarmente i dati**:
  - Lavorare con dati aggiornati per evitare che i modelli si basino su schemi storici non più rilevanti.

4 Esempi di best pratices



Esempi aziendali concreti:

→ **LinkedIn** ha investito in sistemi di intelligenza artificiale per identificare possibili pregiudizi nei dati relativi alle raccomandazioni di lavoro, adottando strategie per garantire che i suggerimenti di carriera siano neutri e basati sulle competenze.

→ **Google** ha adottato un framework interno per verificare l'equità degli algoritmi utilizzati nei suoi servizi, con focus su selezione e valutazione.

5 Implicazioni etiche e legali nell'utilizzo di AI nei luoghi di lavoro

L'intelligenza artificiale sta ridefinendo i processi decisionali nei luoghi di lavoro, ma il suo utilizzo solleva sfide etiche e legali di fondamentale importanza. Per garantire un'applicazione responsabile e giusta, le organizzazioni devono affrontare aspetti legati ai diritti dei lavoratori, alla trasparenza, alla protezione dei dati e alla responsabilità.

Responsabilità delle aziende e protezione dei diritti dei lavoratori

- **Responsabilità nelle decisioni algoritmiche**: le aziende che utilizzano sistemi di IA devono garantire che le decisioni prese siano giuste e non discriminatorie. Questo include la selezione del personale, la valutazione delle prestazioni e la gestione dei dipendenti. Ad esempio, l'uso di algoritmi per licenziare dipendenti, come accaduto con Amazon nei suoi centri logistici, ha sollevato dubbi sull'automazione delle decisioni senza supervisione umana.
- **Tutela contro le discriminazioni**
- **Coinvolgimento dei lavoratori**: è essenziale coinvolgere i dipendenti e i rappresentanti sindacali nella progettazione e implementazione di sistemi di IA per garantire che i loro diritti siano tutelati.

Giustizia e trasparenza nei sistemi di selezione e valutazione

Accesso alle motivazioni delle decisioni

- I lavoratori e i candidati devono avere il diritto di comprendere come e perché una determinata decisione (ad esempio, l'esclusione da una selezione o una valutazione negativa) è stata presa da un algoritmo.

Fairness by Design

- Le aziende devono progettare i loro sistemi con principi di equità integrati fin dall'inizio, considerando l'impatto delle decisioni sugli utenti finali.

Prevenzione di "effetti cascata"

- Una decisione algoritmica ingiusta in una fase (ad esempio, la selezione) può influenzare tutte le fasi successive, come la promozione o il trattamento economico. È fondamentale monitorare l'intero processo.

Normative privacy e protezione dei dati

- I lavoratori devono essere informati su come i loro dati vengono raccolti e utilizzati.
- Devono essere richiesti solo i dati strettamente necessari (principio di minimizzazione).
- È necessario garantire la sicurezza dei dati per evitare abusi.

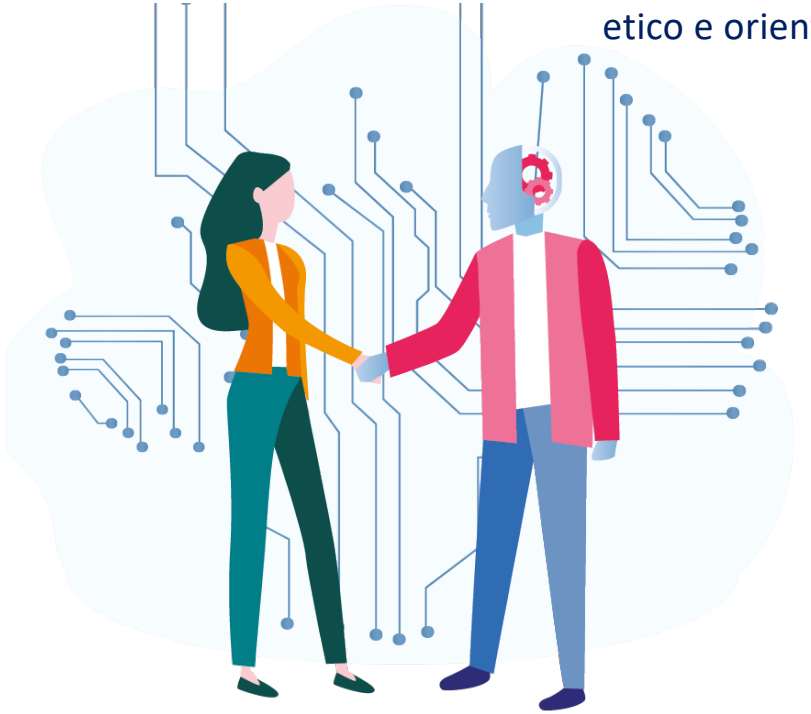
Tutela della privacy

Accountability delle aziende

Le organizzazioni devono documentare e giustificare l'uso dei sistemi di IA, dimostrando di rispettare normative sulla protezione dei dati e principi etici.

6 Necessità dell’intervento umano

Nonostante i progressi nell’intelligenza artificiale (IA), l’intervento umano rimane essenziale nei processi decisionali, specialmente in contesti sensibili come la gestione delle risorse umane. La combinazione tra automazione e supervisione umana garantisce che i sistemi di IA siano utilizzati in modo responsabile, etico e orientato al benessere dei lavoratori.



- **Esempi di casi in cui è necessaria la supervisione umana**: selezione dei candidati, sanzioni disciplinari, promozioni e avanzamenti di carriera;
- **Interpretazione dei risultati e dei dati algoritmici;**
- **Mediazione dei conflitti, empatia e connessione umana.**

Il ruolo dei professionisti HR

I professionisti delle risorse umane svolgono un ruolo cruciale nel bilanciare l'uso dell'IA con l'intervento umano. Ecco come possono integrarsi nel processo:

- Supervisori e validatori delle decisioni dell'IA: gli HR devono verificare che i risultati prodotti dagli algoritmi siano coerenti con gli obiettivi aziendali e privi di bias. Ad esempio, possono revisionare i punteggi assegnati ai candidati dall'IA e prendere decisioni finali basate su una visione più ampia.
- Custodi dell'etica aziendale: gli HR devono garantire che i sistemi di IA siano utilizzati in modo etico e conforme alle normative, come il GDPR in Europa.
- Facilitatori della formazione continua: devono promuovere la formazione del personale su come interagire con i sistemi di IA, rendendo i dipendenti più consapevoli delle potenzialità e dei limiti della tecnologia.