Da dove viene il rischio?

Nel suo studio sull'Intelligenza Artificiale[1], il Servizio di Ricerca del Parlamento Europeo ha affermato che: "*È importante notare che gli algoritmi di IA non possono essere oggettivi perché, proprio come le persone, nel corso della loro formazione sviluppano un modo di dare un senso a ciò che hanno visto in precedenza, e usano questa 'visione del mondo' per categorizzare le nuove situazioni che vengono loro presentate.*" [deepl translation]

Vediamo da dove deriva la soggettività di un'IA e quali sono i rischi associati.

I pregiudizi nei dati e negli algoritmi

Come per qualsiasi sistema digitale, i dati utilizzati nelle piattaforme basate sull'IA provengono da fonti diverse e hanno formati multipli. Sono portatori di diversi tipi di distorsioni[2]. Le distorsioni dei dati sono principalmente di tipo statistico. Ne elenchiamo alcuni.

- La **bias dei dati** è tipicamente presente nei valori dei dati. Ad esempio, un algoritmo di reclutamento addestrato su un database in cui gli uomini sono sovrarappresentati escluderà le donne.
- Lo **stereotype bias** è la tendenza ad agire in riferimento al gruppo sociale di appartenenza. Ad esempio, uno studio dimostra che le donne tendono a cliccare sulle offerte di lavoro che ritengono più facili da ottenere in quanto donne.
- L'**bias delle variabili omesse** (bias di modellazione o di codifica) è un bias dovuto alla difficoltà di rappresentare o codificare un fattore nei dati. Ad esempio, poiché è difficile trovare criteri concreti per misurare l'intelligenza emotiva, questa dimensione è assente dagli algoritmi di reclutamento.
- Il **bias di selezione** è a sua volta dovuto alle caratteristiche del campione selezionato per trarre conclusioni. Ad esempio, una banca utilizzerà i dati interni per ricavare un punteggio di credito, concentrandosi su coloro che hanno o non hanno ottenuto un prestito, ma ignorando coloro che non hanno mai avuto bisogno di un prestito, ecc.

Il pregiudizio algoritmico è principalmente una questione di ragionamento. Tali pregiudizi sono introdotti dagli ingegneri dell'IA deliberatamente o meno.

Il già citato studio del Servizio di Ricerca del Parlamento Europeo fornisce due esempi concreti: "*Si consideri un algoritmo di IA simbolico per l'esame delle domande di lavoro. Potrebbe valutare i candidati assegnando punteggi solo sulla base della loro istruzione ed esperienza. Tuttavia, se non tiene conto di fattori come il congedo di maternità o non riconosce*

in modo appropriato l'istruzione in istituzioni straniere come farebbero i comitati di selezione umani, l'algoritmo potrebbe discriminare le donne e i candidati stranieri ". [deepl translation]

"*Ora, consideriamo uno strumento di IA simile all'interno del paradigma ML (Machine Learning). Tali algoritmi trovano il loro modo di identificare quali tipi di candidati sono stati selezionati nei loro dati di addestramento. Se esiste una storia di pregiudizi strutturali in queste selezioni, ad esempio la discriminazione razziale, l'algoritmo può impararli. Anche quando i dati relativi alla nazionalità o all'etnia vengono rimossi dai dati, l'algoritmo ML è abile nel trovare proxy per i modelli sottostanti in altri dati, come le lingue, i codici postali o le scuole, che possono essere buoni predittori dell'etnia*". [deepl translation]

I tre aspetti del rischio algoritmico

Il rischio algoritmico può essere caratterizzato in tre modi[3].

- In primo luogo, c'è il **confinamento algoritmico**, che può riguardare anche le opinioni, le conoscenze culturali o persino le pratiche commerciali. Infatti, gli algoritmi mettono l'utente di Internet di fronte agli stessi contenuti, a seconda del suo profilo e dei parametri integrati, nonostante il rispetto del principio di equità. È il caso dei siti di raccomandazione di notizie come Facebook o di prodotti come Amazon.
- Il secondo aspetto del rischio algoritmico è legato al **controllo di tutti gli aspetti della vita di un individuo**, dalla regolamentazione delle informazioni per gli investitori alle sue abitudini alimentari, agli hobby o persino allo stato di salute. Questo tracciamento dell'individuo suggerisce una forma di sorveglianza che contravviene all'essenza stessa della libertà individuale.
- Il terzo è legato alla **potenziale violazione dei diritti fondamentali**. In particolare, la discriminazione algoritmica, definita come un trattamento sfavorevole o diseguale, rispetto ad altre persone o ad altre situazioni uguali o simili, basato su un motivo espressamente vietato dalla legge. Ciò comprende lo studio dell'equità (*fairness*) degli algoritmi di classificazione (ordinamento delle persone che cercano lavoro online), di raccomandazione e di apprendimento predittivo. Il problema dei pregiudizi discriminatori indotti dagli algoritmi riguarda diversi settori, come le assunzioni online, le decisioni dei tribunali, le decisioni delle pattuglie di polizia o le ammissioni scolastiche.

Come gestire i rischi legati ai dati e agli algoritmi?

Per R. Schwartz & al.[4], "*Il bias non è né nuovo né unico per l'IA e non è possibile ottenere un rischio zero di bias in un sistema di IA*".
Nel frattempo, riconoscere che gli agenti di IA sono intrinsecamente soggettivi è un prerequisito fondamentale per garantire che vengano applicati solo a compiti per i quali sono ben equipaggiati.

Lo studio dell'EPRS si conclude con diverse raccomandazioni per l'utilizzo di applicazioni basate sull'IA:

- Comprendere i pregiudizi e la soggettività
- Evitare applicazioni che vanno oltre le capacità dell'IA
- Evitare applicazioni con impatti indesiderati
- Mantenere l'autonomia umana
- Cercare soluzioni ai problemi, non problemi per le soluzioni
- Considerare ciò che vogliamo veramente dall'IA

---

1. Artificial intelligence: How does it work, why does it matter, and what can we do about it ? - Philip Boucher, Scientific Foresight Unit (STOA) - ISBN: 978-92-846-6770-3 - Union Européenne, 2020 ↩
2. Algorithms, Data and Bias: Public Policy Needed, Anne Bouverot, Thierry Delaporte, 2019 ↩
3. Article in French: D'où vient le risque ? Des données et des algorithmes - Serge Abiteboul, Thierry Viéville, 2020 ↩
4. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence - Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, NIST Special Publication 1270 , 2022 ↩