Bias di genere e intelligenza artificiale: intervenire ora perché non sia troppo tardi

(di Carmine Andrea Trovato e Giulia Casavola)

Che cosa sono i bias di genere nell'intelligenza artificiale

Prendete il vostro smartphone per sbloccarlo, ma non ci riuscite. Un poliziotto utilizza un software di riconoscimento facciale e scambia un passante innocente per un criminale. Sono errori che i sistemi di facial recognition possono commettere. E possono commetterli con ancora più facilità per i volti delle donne, e tanto più quanto più scura è la loro carnagione.

Uno studio del MIT[1] ha dimostrato che i programmi di analisi facciale adottati da IBM, Microsoft e Amazon sbagliano a identificare i volti femminili con una frequenza del 18% superiore all'incidenza di errore per i volti maschili. La percentuale di errore è particolarmente alta per le donne nere, pari addirittura al 35%, contro lo 0,8% di errore riscontrato per i volti di uomini bianchi.

L'*implicit bias*, quell'insieme di concetti e credenze condivise alla base dei nostri schemi cognitivi, è una condizione connaturata all'essere umano; da un lato è uno strumento di sopravvivenza, che ci consente di prendere decisioni d'intuito, senza dover ponderare di volta in volta ogni azione; dall'altro è terreno fertile per il pregiudizio. Ma se siamo più abituati a identificare fenomeni di razzismo e discriminazione quando sono commessi da esseri umani, lo stesso non si può dire per le applicazioni dell'IA, che vengono anzi dai più percepite come tecnologie neutrali e incapaci di errore. In realtà, i bias umani sono trasferibili all'IA e da questa interiorizzati come errori sistemici, ancora più pericolosi perché amplificati dall'azione su larga scala della macchina. Il gender bias nelle macchine si manifesta in due modi: (i) nei processi decisionali, ad esempio nel *recruiting*; (ii) perpetuando gli stereotipi di genere, com'è il caso degli assistenti virtuali, come Siri, che hanno voci femminili e sono programmate per rispondere con tono remissivo e ammiccante ai messaggi violenti e sessualmente espliciti che gli vengono rivolti dagli utilizzatori.

Questi bias vengono trasferiti all'IA tramite il processo di machine learning, dando in pasto alla macchina set di dati rappresentativi solo di una porzione della popolazione, o la cui selezione riflette un pregiudizio sociale. La macchina sarà quindi incapace di identificare correttamente la parte della popolazione non coperta dalle informazioni ricevute, o riproporrà, nelle sue decisioni, il trend storico che è emerso dalla documentazione analizzata, esasperandolo.

Perchè è fondamentale pensare a un'intelligenza artificiale a prova di bias?

Vista la rapida e pervasiva diffusione della IA[2] degli ultimi anni, è importante intervenire ora perché i bias di genere non costituiscano una minaccia ai diritti fondamentali delle persone e, nella maggioranza dei casi, delle donne. Questo rischio esiste, vista la varietà dei settori in cui l'IA sta iniziando a trovare applicazione: dai sistemi di selezione dei candidati, ai servizi finanziari, ma anche nei motori di ricerca. In tutti questi ambiti, l'IA ha dato prova di aver imparato a riprodurre pregiudizi legati al genere. È il caso dell'algoritmo utilizzato da Amazon per selezionare i candidati migliori per posizioni da software engineer o altri profili tecnici: questo, infatti, si è scoperto avvantaggiare gli uomini, poiché allenato sulla base dei CV dei dipendenti assunti negli anni precedenti, per la maggior parte, appunto, uomini. È interessante anche l'esperienza osservata nell'ambito dei servizi finanziari, in cui può verificarsi che l'IA ritenga con più facilità un uomo idoneo a ricevere un finanziamento rispetto a una donna. Infatti, i documenti analizzati nel processo di machine learning rivelano che storicamente è stato più frequente per le donne vedersi negato un mutuo, ma questo soprattutto in ragione di tutti i limiti che sussistevano fino a pochi anni fa alla possibilità per una donna di finanziarsi da sola, senza l'intermediazione del marito.

Cosa è necessario fare?

Alcuni spunti su come invertire questo trend sono stati proposti dalla Commissione Europea in un report di settembre 2020 sul gender bias[3]. Innanzitutto, il bias nei sistemi di IA deve essere combattuto già dalla fase di progettazione, implementando un approccio di "by design", soprattutto quando l'IA sia destinata a un uso pubblico. Per realizzare questo obiettivo, la *diversity* deve essere perseguita innanzitutto nella formazione dei team di sviluppatori. Sarebbe pertanto utile integrare i percorsi di formazione dei data scientist con corsi multidisciplinari e, in particolare, di etica. Uno step fondamentale riguarda poi la definizione di basi dati attendibili per l'auto-apprendimento dell'IA. Sarebbe infatti inutile formare sviluppatori se poi la macchina continuasse ad apprendere e a determinare le proprie scelte basandosi sull'analisi di informazioni sbagliate o parziali.

Sullo sfondo di questi interventi è necessario che la ricerca continui a progredire, per esplorare le modalità con cui i bias delle macchine impattano le diverse categorie sociali, osservandole dal punto di vista delle intersezioni tra sesso, identità di genere, razza, etnia, orientamento sessuale, e così via.

Conclusioni

L'intelligenza artificiale schiuderà per noi un nuovo mondo, ma questo mondo ha bisogno di regole che mettano l'essere umano al centro – l'ha affermato la presidente della Commissione Europea Ursula Von Der Leyen nel suo discorso sullo stato dell'Unione del 2020[4]. Un punto di partenza cruciale per dare delle regole all'IA è la sua fase di progettazione, che deve, già questa, essere informata al principio di non-discriminazione e coinvolgere, idealmente, sviluppatori consapevoli dei risvolti etici dell'utilizzo dell'IA e rappresentativi di diverse identità sociali. È l'approccio peraltro già adottato dal GDPR con riferimento alla protezione dei dati personali, che, ai sensi dell'art. 25, deve ispirare

la tecnologia fin dal suo concepimento (“privacy by design”). Come? Esistono dei software di bias detection, ad esempio, in grado di individuare gli eventuali pregiudizi sistemici sviluppati dagli algoritmi e di cogliere l’origine dell’errore. Per l’essere umano, d’altronde, è più difficile individuare il bias a posteriori, dopo che la decisione viziata è stata presa.

Non ci si può aspettare di riuscire a evitare del tutto l’errore nella macchina, ma è possibile e doveroso lavorare per ridurne il rischio per imparare a riconoscere quei pregiudizi che diamo per scontati negli esseri umani, ma che tendiamo a sottovalutare nelle macchine.

[1] Disponibile qui: http://gendershades.org/overview.html.

[2] Il mercato globale dell’IA è stato valutato a 39,9 miliardi di dollari nel 2019 e ci si aspetta che cresca a un tasso di crescita annuo composto del 42,2% dal 2020 al 2027. Fonte: https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market.

[3] Disponibile qui: https://ec.europa.eu/info/sites/info/files/research_and_innovation/research_by_area/documents/ec_rtd_gender-bias-in-ai-factsheet.pdf .

[4] Disponibile qui: https://ec.europa.eu/info/sites/info/files/state-of-the-union-speech_it_0.pdf.