Come i pregiudizi dell’intelligenza artificiale rafforzano quelli degli umani

Piccoli pregiudizi, sia umani che derivati dalla programmazione dell’intelligenza artificiale, possono crescere a dismisura fino a creare un effetto valanga. È quanto emerge da un nuovo studio condotto su oltre mille partecipanti da ricercatori dell’University College of London (UCL), sottoposto a revisione paritaria e pubblicato sulla rivista scientifica *Nature Human Behaviour*. Se da una parte l’IA può agevolare e perfezionare il lavoro dell’uomo, dall’altra, quando addestrata principalmente su dati umani, **non solo apprende i bias ed eventuali errori sistematici già presenti, ma li amplifica e li restituisce agli utenti**, influenzando ulteriormente i loro giudizi. Si tratta di un fenomeno che, come spiegato dai ricercatori, rende fondamentale affrontare i “pregiudizi algoritmici” per evitare che piccoli errori si trasformino in **problematiche sistematiche e pervasive**. «È fondamentale che i sistemi di IA vengano perfezionati per essere il più imparziali e accurati possibile», ha commentato il dott. Moshe Glickman, ricercatore dell’UCL e coautore dello studio.

L’analisi ha coinvolto oltre 1.200 partecipanti in una serie di esperimenti, volti a studiare le varie sfaccettature del fenomeno: nel **primo caso**, è stato chiesto a un gruppo di persone di giudicare se i volti in alcune foto apparissero tristi o felici e i dati raccolti, che mostravano una leggera tendenza a giudicare i volti come tristi, sono stati usati per addestrare un algoritmo di intelligenza artificiale. Successivamente, un altro gruppo di partecipanti ha svolto lo stesso compito, ma con il supporto dei suggerimenti forniti dall’IA e, dopo l’interazione, le persone hanno mostrato una tendenza ancora più accentuata nel giudicare i volti principalmente come tristi. Il **secondo esperimento** aveva l’obiettivo di studiare il bias introdotto dagli algoritmi: i partecipanti hanno completato un compito visivo in cui dovevano determinare la direzione di movimento di punti su uno schermo. Sono stati esposti ad un algoritmo accurato, uno rumoroso e uno con un errore sistematico e sono state studiate le tendenze registrate. Infine, il **terzo esperimento** ha coinvolto un sistema di IA generativa ampiamente utilizzato, il quale ha creato immagini di “manager finanziari” **appositamente sovrarappresentando una etnia** rispetto alle altre per poi studiare se i partecipanti erano più propensi ad “aspettarsi” che tale ruolo fosse svolto principalmente da lavoratori dell’etnia sovrastimata.

Lo studio ha evidenziato che l’intelligenza artificiale non solo assorbe i pregiudizi umani, ma li amplifica. I partecipanti del primo esperimento hanno **mostrato un bias ancora più accentuato nel giudicare i volti come tristi**, nel secondo esperimento l’algoritmo accurato ha migliorato la precisione dei giudizi umani mentre quello con errore sistematico **ha influenzato significativamente i partecipanti** e infine, nel terzo esperimento, i sottoposti hanno **evidenziato una inclinazione a scegliere per il ruolo selezionato l’etnia appositamente sovrarappresentata** dall’IA, amplificando così stereotipi già presenti. «Le persone sono intrinsecamente prevenute, quindi quando addestriamo i sistemi

di intelligenza artificiale su set di dati prodotti da persone, gli algoritmi di intelligenza artificiale apprendono i pregiudizi umani incorporati nei dati. L’intelligenza artificiale tende quindi a sfruttare e amplificare questi pregiudizi per migliorare la sua accuratezza di previsione. Abbiamo scoperto che le persone che interagiscono con sistemi di intelligenza artificiale distorti possono diventare a loro volta ancora più distorte, creando un potenziale effetto valanga in cui piccoli pregiudizi nei set di dati originali vengono amplificati dall’intelligenza artificiale, il che aumenta i pregiudizi della persona che utilizza l’intelligenza artificiale», ha commentato la professoressa Tali Sharot, ricercatrice dell’UCL e coautrice dello studio. «Tuttavia, è importante notare che abbiamo anche scoperto che interagire con IA accurate può migliorare i giudizi delle persone, quindi **è fondamentale che i sistemi di IA vengano perfezionati per essere il più imparziali e accurati possibile**. Gli sviluppatori di algoritmi hanno una grande responsabilità nella progettazione di sistemi di intelligenza artificiale; l’influenza dei pregiudizi dell’intelligenza artificiale potrebbe avere implicazioni profonde man mano che l’intelligenza artificiale diventa sempre più diffusa in molti aspetti della nostra vita», ha concluso l’altro coautore, il dott. Moshe Glickman, anche lui ricercatore dell’University College of London.

[di Roberto Demaio]