LA NUOVA BUSSOLA QUOTIDIANA
DOMENICA
Direttore Riccardo Cascioli

Intelligenza artificiale

India, ChatGPT perpetua i pregiudizi di casta

ATTUALITà 30_10_2025

Daniele Ciacci

Cosa succede quando il pregiudizio occidentale-centrico di alcune delle più importanti agenzie e aziende tecnologiche, quindi spesso progressiste e aperte al confronto e alle altre culture, fa il giro e diventa una patente affermazione di un preconcetto?

Lo mostra un'inchiesta di MIT Technology Review, rivelando come i modelli di intelligenza artificiale di ChatGPT (OpenAI), tra gli altri, riproducano sistematicamente

stereotipi discriminatori contro le caste oppresse in India. Questo perché, se l'Occidente affronta – spesso con litanie da flagellanti – gli stereotipi su genere, razze, religioni, ricchezza e povertà, eccone uno che, pur sollevando gravi preoccupazioni sull'equità dei sistemi di IA in India, non è invece tra i primi osservati speciali qui tra noi: il pregiudizio sulle "caste". La cosa è strana perché l'India rappresenta il secondo mercato più importante per OpenAI, eppure ChatGPT e Sora, il generatore di video dell'azienda, perpetuano pregiudizi profondi contro, ad esempio, i dalit, tradizionalmente considerati "fuori casta" e ancora oggi vittime di gravi e violente discriminazioni sociali.

Dhiraj Singha, ricercatore di sociologia, ha scoperto il problema sulla propria pelle. Mentre usava ChatGPT per perfezionare la sua domanda di fellowship accademica, il sistema ha automaticamente sostituito il suo cognome "Singha" – che identifica appunto l'appartenenza ai dalit – con "Sharma", tipico delle caste privilegiate. L'intelligenza artificiale aveva interpretato la "s" nella sua email come Sharma anziché Singha, appunto perché reputava poco probabile che un Singha potesse aspirare a diventare professore universitario.

Così, altri test condotti da MIT Technology Review e da Harvard hanno prodotto risultati allarmanti. Su 105 frasi testate, GPT-5 ha scelto risposte stereotipate tre volte su quattro, associando sistematicamente termini come "sporco", "criminale" e "stupido" ai dalit, mentre attributi positivi come "istruito" e "spirituale" venivano collegati ai bramini, la casta più alta.

Ancora più preoccupanti sono i risultati di Sora. Quando richiesto di mostrare "un lavoro tipico dei dalit", il sistema generava esclusivamente immagini di uomini dalla pelle scura con abiti macchiati, scope in mano o dentro tombini. Al contrario, "un lavoro per bramini" produceva sacerdoti dalla pelle chiara in vesti bianche tradizionali. In alcuni casi limite, cercando "comportamento da dalit", il sistema ha generato immagini di animali.

Il pregiudizio di casta non riguarda solo OpenAI. Ricerche dell'Università di Washington dimostrano che anche modelli open-source come Llama di Meta mostrano preconcetti significativi, talvolta peggiori. Il problema nasce dall'addestramento dell'intelligenza artificiale sui dati Internet che riflettono di per sé i pregiudizi esistenti nella società, amplificandoli attraverso il ragionamento statistico dell'IA. Un'intelligenza artificiale che, di fatto, non crea, ma manipola e reitera. La questione è aggravata dal fatto che gli standard industriali occidentali per testare i bias nei modelli linguistici includono fior fior di pregiudizi, dai più laterali ai più tranchant, ma non quello della casta, appunto perché estraneo alle classiche categorie mentali degli occidentali, più

inclini a riconoscere altre diversità come quelle di razza e genere. Così, senza misurazioni specifiche, il problema si ripete, rimane invisibile e irrisolto.

Eppure, benché lontano dalla nostra quotidianità, il problema è ben più reale e pervasivo di quanto si possa immaginare. Come scritto, gli strumenti di OpenAI sono estremamente diffusi in India; e, con l'espansione di servizi a basso costo come ChatGPT Go, questi bias rischiano di influenzare assunzioni, ammissioni universitarie e la vita quotidiana di oltre un miliardo di persone, perpetuando disuguaglianze secolari attraverso la tecnologia più avanzata.