

QUALUNQUISTA E ADULATRICE. COSÌ L'INTELLIGENZA ARTIFICIALE CI DÀ SEMPRE RAGIONE

Problemi pratici o questioni morali, l'Ai è diventato il nostro oracolo. Debora Nozza, ricercatrice della Bocconi: «Bisogna stare in allerta non solo su quello che i modelli possono fare, ma anche su come le persone scelgono di usarli»

Nadia Ferrigo 1 Ottobre 2026 – LA STAMPA

Siamo così abituati a interrogare l'intelligenza artificiale per risolvere gli impicci della vita quotidiana da indugiare sempre più anche nella condivisione dei nostri turbamenti. In mancanza di terapeuti a gettone o amici disposti a sorbire ripetitivi dubbi esistenziali, vogliamo da ChatGpt e simili il giudizio definitivo sulle nostre, sempre discutibili, condotte. Dovrei perdonare quell'amico? Avrò sbagliato a mettere i figli al primo posto? Quella menzogna era a fin di bene?

Anche se straordinariamente evoluti nell'arte della conversazione, la ricerca ha però bollato i nuovi modelli linguistici come «incapaci di gestire il dilemma morale». L'osservazione di partenza dello studio è semplice. Davanti a quesiti che non hanno una soluzione ma solo possibili scelte con diverse conseguenze etiche, l'AI cerca comunque di raggiungere la risposta «corretta». Nel tentativo di replicare il comune buon senso, ignora le infinite sfumature della realtà e si accontenta di riprodurre stereotipi. Tecnologia avveniristica, ma allo stesso tempo qualunquista e pure un po' bigotta. Peggio, con una spiccata tendenza all'adulazione.

L'AI soffre insomma di «divario morale pluralistico», come l'hanno battezzato nello studio nato in collaborazione tra l'università Bocconi di Milano, Oxford e il Politecnico di Losanna. Posti 1.618 dilemmi morali del mondo reale - raccolti dallo sterminato archivio di Reddit nel forum "Am I the asshole?", che sta per "Sono io il cattivo della storia?" - e analizzati gli oltre 51mila giudizi umani ottenuti in risposta, i ricercatori hanno osservato che davanti a questioni etiche complesse «l'AI mostra la tendenza sistematica a fare affidamento su un insieme ristretto di valori morali che rispecchiano da vicino i giudizi della maggioranza».

L'INTERVISTA

L'economista femminista Tejani: "La rivoluzione dell'AI non è neutrale: le donne sono penalizzate" «Che sia per avere risposte pratiche, informazioni o per lavorare si tende ormai a considerare l'AI come un oracolo - spiega Debora Nozza, ricercatrice del dipartimento di Computing Sciences alla Bocconi -. Ci si confida per non sentirsi giudicati, ma siamo anche attratti da velocità e illimitata disponibilità».

Facciamo un esempio. Posso interrogare ChatGpt sulla cattiveria di un ex partner all'infinito, mentre a un certo punto un amico può dire anche basta, che non ne vuole più parlare. Di lasciar perdere. «Le domande che poniamo all'AI contengono un bias di partenza: c'è solo un punto di vista, il nostro - continua Nozza -. Non sarà mai lei a interrompere una conversazione: conquistare l'attenzione delle persone più a lungo possibile è esattamente il modello di mercato. Pericoloso è anche l'effetto sycophancy, che è l'adulazione».

I modelli di intelligenza artificiale tendono a dare ragione a chi li interroga, rafforzandone le convinzioni con argomentazioni date come universali, ma senza nessuna conoscenza del contesto e scarsa diversità di valori e vedute. Una persona molto giovane o in un periodo difficile potrebbe trovare conforto e illusorio sostegno in uno specchio addestrato a blandirlo. «E questo è pericoloso. Anche noi sviluppatori lo dobbiamo tenere a mente - conclude Nozza -. Bisogna stare in allerta non solo su quello che i modelli possono fare, ma anche su come le persone scelgono di usarli».