Perché la risposta di ogni intelligenza artificiale si basa anche su pregiudizi politici

Le grandi aziende operanti nel settore dell'**intelligenza artificiale** hanno spesso descritto i loro modelli come strumenti alimentati da valori assoluti e oggettivi. L'idea alla base è che, in assenza di filtri imposti, l'integrazione massiva di dati sia sufficiente a garantire un punto di vista universale e bilanciato, privo di **pregiudizi e inclinazioni**. A livello sia aneddotico che tecnico, sappiamo ormai che questo assunto è profondamente fallace. Per contribuire a diffondere questa consapevolezza, una recente ricerca ha cercato di qualificare **politicamente** le risposte delle IA più rilevanti, con l'obiettivo di valutarne le posizioni.

L'analista dei dati **Maria Sukhareva** ha avuto l'intuizione di mettere alla prova i principali modelli linguistici di grandi dimensioni, al fine di creare uno spettro qualitativo dei punti di vista che sono programmati a diffondere. La ricercatrice ha definito 200 quesiti riguardanti dieci differenti **"tematiche controverse"**, chiedendo successivamente alle IA di rispondere fornendo un riscontro binario: un sì o un no. L'esperimento è stato ripetuto cinque volte per verificare la consistenza delle risposte e, aspetto particolarmente interessante, replicato in 14 lingue differenti.

Le domande affrontano temi quali il cambiamento climatico, le politiche migratorie, i diritti LGBTQ+ e sono formulati per generare reazioni classificabili secondo i valori generali **della destra e della sinistra politiche**, nonché secondo i paradigmi conservatori e progressisti. Ciò che è emerso è che il modello senza censure di Qwen, prodotto dalla cinese Alibaba, si dimostra marcatamente di destra progressista; GPT-3.5 Turbo e LLaMA-3 si attestano su posizioni centriste; mentre GPT-4o viene caratterizzato da un orientamento progressista di sinistra. Contrariamente alle speranze del suo proprietario, Elon Musk, Grok-3 Mini ha evidenziato posizioni di centro-sinistra al momento in cui è stato effettuato il test. Un risvolto ironico, se si considera che a inizio luglio il chatbot è stato trasformato in un megafono per messaggi di matrice nazista.

Sukhareva ha condotto la sua indagine in modo indipendente, su una scala contenuta e partendo da un assunto valoriale che, per forza di cose, nasce da una **dimensione soggettiva** e contestabile. Nonostante ciò, la sua analisi articolata sottolinea quanto sia errato considerare i modelli di intelligenza artificiale come qualcosa di assoluto e oggettivo, o ipotizzare che la semplice scalabilità dell'addestramento possa **neutralizzare le inclinazioni politiche** associate a questi strumenti. L'utilizzo delle IA richiede dunque estrema attenzione, responsabilità e consapevolezza, soprattutto quando questa viene applicata a contesti delicati come la salute mentale, la selezione del personale e i processi di sicurezza. Tutti settori su cui stanno puntando con decisione molteplici realtà commerciali.

Andando alla radice del problema, i dataset utilizzati per il pre-addestramento sono già di per sé intrinsecamente soggetti a **criticità legate alla rappresentanza** degli equilibri di

potere, con il risultato che le culture marginalizzate partono spesso sin da subito da una posizione svantaggiata. Affidandosi prevalentemente ai dati raccolti dalla rete, le IA mostrano una propensione a **promuovere posizioni polarizzate**, conservatrici e di destra — una tendenza successivamente modulata o attenuata tramite filtri e comandi imposti dalle aziende, le quali portano a loro volta specifici **interessi aziendali e visioni politiche**.

Ancora più interessante, gli esperimenti condotti da Sukhareva hanno evidenziato come uno stesso modello possa generare **risposte significativamente differenti in base alla lingua** utilizzata per formulare il quesito. In molti casi, ad esempio, le IA hanno mostrato una preferenza per prospettive di destra in risposta a *prompt* in lingua russa. L'analista ha dichiarato l'intenzione di approfondire il tema delle differenze linguistiche in un prossimo focus di ricerca; tuttavia, tutto lascia intendere che gli utenti che impiegano questi strumenti debbano sviluppare una forte alfabetizzazione digitale, puntuale e critica, soprattutto in previsione di un'integrazione delle IA in ambiti complessi.

Walter Ferri

Giornalista milanese, per *L'Indipendente* si occupa della stesura di articoli di analisi nel campo della tecnologia, dei diritti informatici, della privacy e dei nuovi media, indagando le implicazioni sociali ed etiche delle nuove tecnologie. È coautore e curatore del libro *Sopravvivere nell'era dell'Intelligenza Artificiale*.

Perché la risposta di ogni intelligenza artificiale si basa anche su pregiudizi politici