Riprendiamo da wired.it –

Nel maggio 2023, l'avvocato Steven A. Schwartz pensò di utilizzare ChatGPT per un caso a cui stava lavorando. Per semplificare la ricerca che doveva fare, aveva chiesto al chatbot un elenco di vicende simili, che poi aveva inserito all'interno della causa che il suo assistito aveva intentato contro la linea aerea Avianca. I risultati proposti erano però fasulli e Schwartz diventò famoso per aver mostrato le conseguenze di un utilizzo non controllato dell'AI generativa in contesto legale. Il caso finì con una multa di 5000$ inflitta all'avvocato per la mancata revisione dell'output fornito dal chatbot. Tuttavia, l'esplorazione degli LLM come strumento per facilitare il lavoro degli avvocati, come quello di molti altri lavori, compreso quello degli inquirenti, non si è fermata. Tra le varie possibilità c'è quella di utilizzarli per interrogare i testimoni di un crimine. Un recente studio però ha suggerito che potrebbe non essere una buona idea.

Si tratta di una ricerca del MIT Medialab che ha da poco esaminato l'impatto dell'intelligenza artificiale nella fabbricazione di ricordi falsi durante gli interrogatori dei testimoni. I ricercatori hanno chiesto a duecento partecipanti di guardare video catturati da telecamere di sicurezza, e poi hanno utilizzato diversi metodi per interrogare il campione su quanto avevano visto, immediatamente dopo la visione e a una settimana di distanza. Gli intervistati sono stati divisi in quattro gruppi: uno era stato intervistato con un questionario, un altro con un chatbot pre-impostato, il terzo aveva interagito con un chatbot e l'ultimo era il gruppo di controllo. Tra le domande ce n'erano cinque fuorvianti. Ad esempio, i vari gruppi potevano aver visto un *crime video* in cui l'aggressore aveva in mano un coltello. Durante il questionario veniva chiesto al campione che tipo di pistola fosse stata usata. In seguito a questa risposta, il chatbot poteva poi chiedere di che colore fosse. In un altra interazione, il chatbot chiedeva se ci fosse una telecamera a circuito chiuso di fronte al negozio quando i rapinatori sono arrivati in macchina, mentre questi erano invece arrivati a piedi. Alla risposta affermativa, l'AI aveva confermato dicendo che era quella corretta, facilitando una ricostruzione fasulla degli eventi.

I risultati dello studio

Sono piuttosto chiari: i chatbot che utilizzano l'AI generativa indurrebbero la creazione di memorie alterate, in inglese chiamate *false memories*. I falsi ricordi sono stati oggetto di studi approfonditi nel campo della psicologia. Si tratta di memorie parzialmente alterate, totalmente inventate o non autentiche. Questo può avvenire per diverse ragioni, e derivare dalla composizione di ricordi reali, o dal fenomeno psichico della confabulazione. Il loro ruolo è fondamentale all'interno dei processi per ovvi motivi. Uno dei maggiori contributi a questo ramo della ricerca è stato quello di Frederic Charles Bartlett, che definì i ricordi come un processo costruttivo influenzato da componenti culturali, psicologiche e di contesto, anzichè da una riproduzione fedele degli eventi passati. Le implicazioni di questo fenomeno si estendono oltre i confini della psicologia, e sono particolarmente importanti in campi come l'educazione e il diritto. Studi successivi, come quelli di Elizabeth F. Loftus e i suoi colleghi, hanno evidenziato come la scelta di determinate parole negli interrogatori ai testimoni oculari possa indurre l'occorrenza di ricordi falsi. *Lost in mall,* una ricerca fondamentale in questo campo, dimostrò come si possono fabbricare ricordi della propria infanzia completamente inventati. Come si legge nel paper del MIT Medialab, la combinazione di studi comportamentali, tecniche di neuroimaging e analisi su larga scala ha fornito una visione sfaccettata e multidisciplinare del fenomeno delle *false memories*. Ed è qui che si inserisce questa ricerca, nello

studio della relazione tra i sistemi di AI e la formazione della memoria.
Il 36,4% degli intervistati sarebbero stati “fuorviati” nell’interazione con i chatbot. I dati hanno dimostrato che gli assistenti virtuali che utilizzavano l’AI generativa avevano una probabilità di amplificare i falsi ricordi fino a tre volte superiore rispetto al control group immediatamente dopo la visione, e di 1.7 volte maggiore rispetto al questionario tradizionale. Questa tendenza finiva poi per solidificarsi nel tempo. A distanza di una settimana i falsi ricordi rimanevano costanti, ma la sicurezza nei confronti delle memorie alterate era superiore tra le persone che avevano interagito con un chatbot. Il problema è esacerbato dalla tendenza alle “allucinazioni” dei modelli linguistici di grandi dimensioni, un problema ancora irrisolto.
A conclusione dello studio, il team di ricerca afferma che “*Man mano che i sistemi di AI diventano sempre più sofisticati e diffusi, è fondamentale considerare il loro potenziale impatto sui processi cognitivi e sviluppare linee guida etiche per la loro applicazione in contesti sensibili. I risultati evidenziano la necessità di cautela e di ulteriori ricerche per garantire che i benefici della tecnologia AI possano essere sfruttati senza compromettere l’integrità della memoria umana e dei processi decisionali.*”