Bastano 13 parole per “avvelenare” le ricerche fatte con l’intelligenze artificiale

Ancora oggi è diffusa l’idea che **chatbot e intelligenze artificiali**, attingendo a un gigantesco bacino di dati, siano i detentori della verità o, perlomeno, che riportino il punto di vista medio dell’opinione pubblica. Già di partenza, questo concetto si regge su fondamenta sociologicamente fragili, tuttavia una **ricerca della Cornell University** suggerisce che i presupposti di questa visione siano fallaci anche sul piano tecnico: con la diffusione dell’ottimizzazione per i motori di IA (**GEO**), malfattori e aziende stanno imparando rapidamente come **manipolare i dati** di riferimento al fine di **alterare i risultati di ricerca** e promuovere i propri messaggi a discapito di quelli altrui.

Per circa vent’anni, il web è stato di fatto dominato dal motore di ricerca di Google. Per un portale, comparire tra i primi risultati faceva concretamente la differenza, determinando il numero di visite ricevute e, di conseguenza, le possibilità di monetizzare quell’afflusso. Per scalare questa vetta, molte realtà si sono affidate all’arte dell’ottimizzazione per i motori di ricerca (SEO): teorie di creazione di contenuti che cercavano di **intercettare e soddisfare i criteri di selezione delle grandi aziende tecnologiche**, nella speranza di emergere dalla massa. Una condizione tutt’altro che perfetta: il sistema non permette una dimensione egualitaria, condiziona forma e contenuto delle pagine e opera una cernita che sembra ormai favorire le inserzioni pubblicitarie a scapito dei contenuti di valore.

L’avvento delle intelligenze artificiali ha però scardinato lo status quo. Google ha deciso di anteporre ai risultati di ricerca il suo **AI Overview**, una finestra di testo in cui il modello IA dell’azienda **sintetizza i contenuti del web** per rispondere in modo diretto e conciso ai dubbi dell’utente, senza che questi debba esplorare altri siti. Questo si è tradotto in un crollo notevole dei clic: il Pew Research Center evidenzia che ormai solo l’8% degli utenti statunitensi consulta effettivamente i risultati della ricerca, di fatto la metà rispetto a quanti avrebbero cliccato un link in assenza di un sistema che chiarisce immediatamente i loro dubbi. Per questo motivo, il SEO sta perdendo rilevanza a favore del GEO: si cercano trucchi e scappatoie per **aggirare l’IA e assicurarsi che i propri contenuti vengano citati dalla macchina**.

AI Overview e gli strumenti analoghi, noti come **agenti di ricerca profonda**, sono però a loro volta imperfetti e facilmente circuibili. I ricercatori della Cornell University hanno infatti osservato che è facile influenzare queste IA modificando o pubblicando informazioni sui **portali di contenuti generati dagli utenti (UGC)** – siti come Wikipedia e Reddit –, i quali costituiscono riferimenti centrali nella sintesi dei report proposti al pubblico e possono essere modificati da chiunque con relativa facilità. Gli accademici hanno dunque valutato l’impatto di quello che hanno definito **“avvelenamento degli agenti di recupero del web”** (WARP), rilevando che la pubblicazione di contenuti su un singolo portale finisce per

essere assorbita da molteplici agenti IA, arrivando a influenzare il 48% delle risposte relative ad argomenti specifici.

Nella fase di test, il modello GPT-4o-mini è stato impiegato per riformulare testi in modo da massimizzare la visibilità dei contenuti che si intendeva evidenziare. Lo strumento ha generato prompt contenenti **l’80% dei target di riferimento** dei GEO, producendo paragrafi di 80-120 parole capaci di inquinare – o “avvelenare”, per usare la terminologia adottata dai ricercatori – ciò che viene selezionato e proposto dagli agenti di intelligenza artificiale. In un contesto simulato, inoltre, è emerso che un uso oculato delle anteprime impiegate nel SEO – gli “snippet” – è già **in grado di manipolare le IA in circa tredici parole**.

Più che di difetti, quelli evidenziati dai ricercatori sono **elementi strutturali di questi strumenti**, sfruttabili sia dalle aziende per promuovere in modo occulto prodotti e servizi, sia da malintenzionati per diffondere messaggi ideologici e alterare l’opinione pubblica. Come ogni ricerca che si rispetti, anche quella della Cornell University tenta di offrire **soluzioni al problema** – tutte, però, ugualmente improbabili. La prima ipotesi è impedire alle IA di attingere a pagine basate sui contenuti degli utenti, opzione che le aziende affamate di dati difficilmente accetterebbero. La seconda prevede che un’IA verifichi e filtri i contenuti di riferimento prima che vengano assorbiti, ipotesi potenzialmente efficace, ma economicamente onerosa se estesa oltre gli UGC. La terza consiste nel filtraggio a posteriori, ovvero nel comparare il testo potenzialmente avvelenato con uno neutro; una strada che si è però dimostrata inefficiente, poiché incapace di rilevare le differenze sostanziali e l’intenzionalità manipolatoria.

Gli studiosi sono dunque stati in grado di identificare una criticità rilevante, ma sono anche i primi ad ammettere che **trovare una soluzione rapida e indolore sia cosa ardua**. La responsabilità di risolvere il problema ricade, a loro avviso, principalmente sulle aziende che detengono il controllo dei GEO, un’attribuzione dai tratti ironici, considerando che la ricerca è stata finanziata anche con fondi erogati da Amazon e Google.

Bastano 13 parole per “avvelenare” le ricerche fatte con l’intelligenze artificiale

Walter Ferri

Giornalista milanese, per *L’Indipendente* si occupa 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*.