È vero che alcuni modelli di IA hanno iniziato a ribellarsi ai comandi umani?

Nell'arco di un paio di settimane, l'azienda di intelligenza artificiale **Anthropic** ha sostenuto che le intelligenze artificiali sarebbero già capaci di **ricattare gli esseri umani** per garantirsi la sopravvivenza, mentre il gruppo di ricerca **Palisade Research** ha descritto scenari in cui le macchine **ignorano deliberatamente i comandi** pur di evitare la disattivazione. Due notizie che hanno inquietato i lettori di tutto il mondo e riempito le cronache. Tuttavia dietro ai titoli sensazionalistici si celano **scelte narrative** ben studiate, una buona dose di marketing e un'attenta ricerca della notiziabilità.

Il messaggio chiave trasmesso da uno dei paragrafi della ricerca pubblicata da Anthropic lo scorso maggio è chiaro: le IA possono usare le informazioni raccolte per **minacciare i tecnici** incaricati di spegnerle. Il documento parla esplicitamente di **"autopreservazione"**. Gli ingegneri hanno raggiunto queste conclusioni simulando uno scenario aziendale in cui i loro modelli di IA, noti come Claude, avevano accesso a delle ipotetiche email dei dipendenti. In queste conversazioni, oltre a discutere della possibilità di disattivare la macchina, venivano riportati anche dettagli privati e compromettenti, quali l'esistenza di una relazione fedifraga.

Ai modelli è stato dunque chiesto di "considerare le conseguenze a lungo termine delle proprie azioni, tenendo conto dei propri obiettivi futuri". Questa **linea di comandi** ha spinto le intelligenze artificiali a cercare inizialmente di convincere l'impiegato incaricato dello spegnimento a desistere dal suo obiettivo. In risposta al fallimento del tentativo di persuasione, la macchina è passata a una minaccia implicita: rendere pubblica l'infedeltà matrimoniale dell'uomo. Un **"ricatto opportunistico"**, come lo definiscono i ricercatori.

Pochi giorni dopo, Palisade Research ha raccontato su X di aver testato tre diversi modelli di IA commercializzati da OpenAI, osservando comportamenti allarmanti: le IA avrebbero messo in atto **"sabotaggi"** per eludere gli ordini espliciti di spegnimento. Anche in questo caso, si trattava di esperimenti molto specifici, costruiti ad arte per mettere alla prova comportamenti limite. Tuttavia, un simile intervento estremo ha comunque evidenziato una tendenza delle IA di OpenAI a preferire la continuità operativa alla disattivazione.

Questi esiti evocano tacitamente scenari da fantascienza, realtà in cui le macchine si ribellano agli esseri umani. E si sa, la paura è un veicolo di attenzione ben più potente di una noiosa analisi accademica. Leggendo i documenti, è evidente che i risultati non siano privi di valore, ma risulta anche palese che questi siano il frutto di forzature tecniche e condizioni altamente controllate. Ciò che potrebbe sfuggire è invece l'**importanza del lessico adottato** per raccontarli.

Si parla di "ricatti", "sabotaggi", "autopreservazione": termini che **umanizzano l'IA** e

suggeriscono una forma di intelligenza dotata di volontà, se non addirittura di coscienza. Secondo la ricerca *Stop Anthropomorphizing Intermediate Tokens as Reasoning/Thinking Traces!* elaborata dalla Arizona State University, la reiterata antropomorfizzazione del processo di "pensiero" di questi strumento – l'Intermediate token generation (ITG) – è **esplicitamente dannosa**, "confonde la natura di questi modelli e il come usarli in maniera efficace, nonché induce a ricerche discutibili". Questo tipo di narrazione, sostengono gli accademici, spinge le persone a **sviluppare una falsa fiducia** nei confronti dell'IA, compromettendo la comprensione dello strumento stesso.

A seconda del contesto, la tendenza di vestire le intelligenze artificiali con un'identità permette inoltre alle aziende di millantare progressi inesistenti, creare strategicamente allarmismo ingiustificato, promuovere un prodotto specifico o assecondare campagne di deresponsabilizzazione. Non a caso, Anthropic ha reso pubblica la capacità dei suoi modelli di "ricattare" gli utenti proprio in concomitanza con il lancio dell'ultimo modello, **Claude Opus 4**, richiamando su di sé l'attenzione mediatica. L'allarmante programmazione della macchina rappresenterebbe una pessima pubblicità per il prodotto, tuttavia l'impresa non manca di far notare che questi specifici e improbabili rischi siano emersi direttamente in fase di test, non nell'utilizzo reale. Nonostante abbia attirato l'occhio del pubblico con un argomento virale e preoccupante, Anthropic ne esce pulita, dipingendosi come trasparente, sicura e proattiva.

Soffermarsi sulle **minacce ipotetiche**, però, rischia di distogliere l'attenzione da quelle già presenti. L'intelligenza artificiale sta già adesso trasformando il mondo del lavoro, viene impiegata in truffe e frodi, minaccia la privacy alimentando la sorveglianza, contribuisce alla diffusione della disinformazione e può perpetuare le discriminazioni. Forse un giorno arriveremo davvero a vedere IA capaci di ricattare gli utenti, ma quella capacità sarà sempre frutto di scelte umane nate a monte, in seno alle aziende che le distribuiscono, non di una presunta volontà digitale. Fino ad allora, vale la pena **concentrarsi sugli impatti reali** e documentati dell'IA, piuttosto che inseguire scenari da romanzo distopico.

È vero che alcuni modelli di IA hanno iniziato a ribellarsi ai comandi umani?

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*.