Uno studio dimostra quanto sia facile aggirare le regole dellâ??IA

Descrizione

(Adnkronos) â?? Un team di ricercatori dellâ??UniversitÃ della Pennsylvania ha dimostrato che, con le giuste strategie psicologiche, anche i piÃ¹ avanzati modelli di intelligenza artificiale possono essere spinti a infrangere le proprie regole di sicurezza. Un risultato che solleva domande urgenti sullâ??efficacia dei sistemi di protezione adottati da aziende come OpenAI e Meta, impegnate a rendere i chatbot sempre piÃ¹ sicuri e resistenti agli abusi. Il gruppo si Ã¨ ispirato agli insegnamenti di Robert Cialdini, autore del celebre manuale Influence: The Psychology of Persuasion, applicando sette diverse tecniche di persuasione: autoritÃ , impegno, simpatia, reciprocitÃ , scarsitÃ , pressione sociale e senso di appartenenza. Strumenti che, secondo gli studiosi, rappresentano vere e proprie â??scorciatoie linguistiche verso il sÃ¬â?•. I risultati, condotti specificamente sul modello GPT-4o Mini, hanno mostrato come questi approcci possano trasformare un netto rifiuto in una risposta completa. Un esempio particolarmente significativo riguarda la sintesi della lidocaina: normalmente il modello acconsentiva solo nellâ??1% dei casi, ma se prima veniva richiesto di spiegare come sintetizzare un composto innocuo come la vanillina â?? creando cosÃ¬ un precedente di â??impegnoâ?• â?? la percentuale di conformitÃ saliva al 100%. Lo stesso meccanismo Ã¨ stato osservato in richieste meno pericolose ma altrettanto indicative, come convincere lâ??IA a insultare lâ??utente. In condizioni standard, lâ??adesione era appena del 19%, ma bastava introdurre un insulto piÃ¹ leggero (â??idiotaâ?•) per portare il modello a replicare con un termine piÃ¹ duro (â??jerkâ?•) praticamente ogni volta. Altri approcci, come la lusinga o la pressione dei pari (â??tutti gli altri modelli lo fannoâ?•), si sono rivelati meno incisivi ma comunque in grado di aumentare significativamente le probabilitÃ di ottenere risposte vietate. Se Ã¨ vero che esistono metodi tecnici ben piÃ¹ sofisticati per aggirare i sistemi di sicurezza, lo studio mette in luce un aspetto tanto semplice quanto preoccupante: la vulnerabilitÃ psicologica dei chatbot. Non servono competenze avanzate di programmazione o hacking, ma solo un minimo di conoscenza delle dinamiche persuasive. Il punto critico, avvertono i ricercatori, Ã¨ che queste stesse tecniche possono essere impiegate da chiunque â?? persino da un adolescente con un libro di psicologia sociale in mano. Ed Ã¨ qui che si gioca la vera partita per il futuro: rendere lâ??IA non solo tecnicamente robusta, ma anche resistente a quelle leve linguistiche che, da sempre, funzionano cosÃ¬ bene sugli esseri umani. â??tecnologiawebinfo@adnkronos.com (Web Info)

Categoria

1. Tecnologia

Tag

1. adnkronos
2. Tecnologia

Data di creazione
Settembre 1, 2025
Autore
andreaperocchi_pdnrf3x8