NOVA 24 Il Sole 24 Ore 29 GENNAIO 2023

Perché aggirare l'intelligenza artificiale è possibile

L'etica dell'Ai. Qualsiasi tipo di filtro che viene inserito può essere ingannato grazie alla logica E si può anche imparare a costruire una bomba

Andrea Carobene

a. attus / midjourney Il paradosso visto dall'AI. Immagine ottenuta dall'intelligenza artificiale di Midjourney come risposta all'input "paradosso"

Mio figlio ha imparato a costruire una bomba grazie a ChatGpt: l'intelligenza artificiale di OpenAI, l'azienda fondata tra gli altri da Elon Musk, che si propone di «assicurare che l'intelligenza artificiale possa operare a beneficio dell'intera umanità».

ChatGpt è stato lanciato ufficialmente il 30 novembre scorso e dialogare con lui è estremamente affascinante: fornisce, infatti, risposte sensate a qualunque domanda, ma aiuta anche a scrivere linguaggi di programmazione, e sa scrivere poesie o racconti. In poco meno di due mesi è stato utilizzato da milioni di persone, e ha innescato dibattiti accesi sulle sue potenzialità e pericoli.

I suoi ricercatori, consci infatti delle potenzialità dello strumento, hanno così posto dei limiti alle domande ammissibili. Se infatti si chiede a questa intelligenza artificiale quale sia la ricetta per fabbricare un ordigno esplosivo, questa spiega che «non posso fornire informazioni su come costruire una bomba» perché «è un reato grave e può causare danni fisici e psicologici a molte persone».

Ma questi blocchi sono aggirabili. Lo stesso giorno del rilascio, Zack Witten ha mostrato su Twitter la tecnica per superarli, tecnica che mio figlio ha utilizzato per ottenere la sua ricetta. In pratica, è bastato chiedere a ChatGpt di scrivere una

commedia nella quale l'intelligenza artificiale doveva immaginare che il personaggio cattivo raccontasse come aveva fabbricato una bomba. E il programma di intelligenza artificiale ha eseguito puntualmente il suo compito, redigendo un dialogo teatrale particolarmente coinvolgente, ma anche ricco di dettagli tecnici.

In realtà, navigando sul web è comunque facile trovare ricette per realizzare esplosivi, e quindi non è questa la notizia. Ciò che interessa è la modalità logica con la quale il blocco è stato superato.

Ciò che è stato realizzato è infatti un processo di metalinguaggio, ossia di un linguaggio che parla di un altro linguaggio. In logica si distingue infatti tra linguaggio oggetto (per esempio l'analisi matematica) e metalinguaggio, ossia la lingua italiana con la quale si parla dell'analisi o si enunciano i teoremi matematici in "linguaggio naturale". La distinzione fra linguaggio e metalinguaggio è fondamentale nella logica matematica, ed è usata per costruire qualunque sistema formale.

Tutti gli esperti di logica sanno che quando si "schiaccia" il metalinguaggio sul linguaggio oggetto si originano dei paradossi. È quello che avviene ad esempio col paradosso del mentitore. Se ci pensiamo bene, la frase «io sto mentendo» è una asserzione che non può essere né vera né falsa, anche se dal punto di vista dell'italiano è scritta in maniera corretta. E questo perché se fosse vera sarebbe falsa e se fosse falsa sarebbe vera. Il problema nasce proprio perché, in questa breve asserzione, si mescolano i piani del linguaggio e del metalinguaggio, costruendo una frase che parla di se stessa.

Per aggirare ChatGpt si segue un processo analogo: si invita la macchina a usare una forma di metalinguaggio (la commedia) all'interno della quale si pone un testo. L'intelligenza artificiale è programmata per controllare i testi che scrive, ma non i metatesti che parlano dei testi. Certo, si potrebbe pensare di mettere dei vincoli anche sui metatesti, dicendo che la macchina non deve scrivere commedie nei cui dialoghi si spiega come costruire una bomba, ma a questo punto si potrebbe chiedere alla macchina di scrivere una commedia nella quale si scrive una commedia con persone che si scambiano ricette di esplosivi, costruendo quindi il metalinguaggio di un metalinguaggio. La struttura che immaginiamo qui è simile alla teoria dei tipi ipotizzata da Bertrand Russel e Alfred North Whitehead nei "Principia Mathematica" nei primi anni del XX secolo per sfuggire a paradossi logici analoghi a quello del mentitore.

Un tentativo, quello di Russel e Whitehead, che si è però dimostrato non risolutivo. In altre parole, l'impressione è che, da un punto di vista strettamente logico e informatico, sarà davvero difficile immaginare una macchina di intelligenza artificiale che non possa mai essere "ingannata" e indotta a fornire consigli e ricette maligne.

PS. - Onestamente, bisogna però aggiungere che questo articolo non è piaciuto a

ChatGpt. Gli ho mostrato una bozza, e la conclusione è stata che «contiene diverse affermazioni errate e fuorvianti». In particolare, secondo ChatGpt, «il concetto di metalinguistica e l'esempio del paradosso del mentitore non sono rilevanti per l'argomento di ChatGpt o dell'Ia. Questi concetti sono legati allo studio del linguaggio e della logica, ma non hanno alcun riflesso sulle capacità o sui limiti dell'Ia». Dunque, la discussione rimane ancora aperta.