I pericoli dell'IA che "pensa"

OpenAI ha ufficialmente lanciato o1, un tipo di Intelligenza artificiale che desta qualche preoccupazione tra gli esperti

Achille Paliotta
Pubblicato 24 Settembre 2024 su Il Sussidiario.net

Lo scorso 12 settembre OpenAI ha ufficialmente lanciato il suo nuovo modello di Intelligenza artificiale (IA), "OpenAI o1". Quest'ultimo rappresenta una decisa evoluzione rispetto ai precedenti, come GPT-4, introducendo innovazioni che migliorano le capacità di ragionamento e problem-solving.

Il modello "o1" utilizza una tecnica chiamata catena di pensieri (*chain-of-thought,* CoT), che consente di migliorare notevolmente la capacità di risolvere problemi complessi, come la risoluzione di equazioni matematiche. Inoltre, in questo modello è stato implementato anche l'apprendimento per rinforzo (*reinforcement learning*) per affinare l'accuratezza e le capacità di ragionamento permettendo al sistema di apprendere dai propri errori, e così migliorare nel corso del tempo, soprattutto in ambiti come la programmazione e i compiti di carattere scientifico. Di conseguenza, invece di fornire immediatamente una risposta, "o1" viene incoraggiato a "pensare", generando una sequenza di passaggi logici che conducono alla risposta finale, grazie al processo di ragionamento incorporato della CoT.

Questo processo permette di "riflettere" su ogni fase e di costruire così una catena logica coerente. Per questa ragione "o1" viene commercializzato come un'IA che "pensa". La funzionalità del monitoraggio dei processi di pensiero latente consente, inoltre, di capire come giunge alle sue conclusioni, rivelando così, almeno in tesi, aree di ulteriore miglioramento. A riprova di ciò i test interni hanno mostrato che il modello ha ottenuto risultati significativamente migliori rispetto a GPT-4.

In linea generale, il modello "pensa" per circa 10 secondi prima di iniziare a scrivere le risposte, e ciò potrebbe essere davvero un tempo limite per moltissimi casi d'uso, così come messo in risalto da diversi sviluppatori. Costi più elevati e tempi di risposta più lenti del modello, dunque, potrebbero fortemente pregiudicare la qualità dell'esperienza dell'utente.

Nondimeno, OpenAI sostiene che "o1" è in grado di superare le prestazioni umane di un dottorando nella risoluzione di problemi specifici, principalmente in campi come biologia, chimica e fisica, dove la precisione è fondamentale. A solo titolo esemplificativo, negli esami di qualificazione per le Olimpiadi internazionali di matematica, il nuovo modello ha risolto correttamente l'83% dei problemi, rispetto a solo il 13% risolto dal suo predecessore, GPT-4o. Ciò lo potrebbe rendere uno strumento prezioso per i ricercatori alle prese con questioni complesse o con grandi quantità di dati.

Il ragionamento a catena consente, inoltre, a "o1" di aderire meglio alle politiche e alle linee guida sulla sicurezza poiché è in grado di ragionare sulla sicurezza nel contesto dato, il che migliora la sua prestazione nei *benchmark* relativi alla generazione di risposte illecite e al *jailbreak*, vale a dire ai tentativi di aggirare i suoi vincoli etici. Questo può essere visto da alcuni come un vincolo non

necessario, ma le imprese che devono affrontare le conseguenze e le responsabilità di ciò che fa un'IA in loro nome potrebbero preferire un modello più sicuro che, ad esempio, non avalli pulsioni suicide, che non produca contenuti illegali, che è meno incline a proporre o accettare conversazioni che potrebbero comportare perdite finanziarie, ecc. Infine, "o1" è particolarmente performante quando si tratta di *coding*. Nelle competizioni Codeforces, difatti, il modello ha raggiunto l'89° percentile, un risultato di tutto rilievo. Dal *debug* del codice in tempo reale alla ricerca scientifica, quindi, "o1" sembra potersi adattare a un'ampia gamma di applicazioni professionali.

Questa decisa evoluzione di "o1", tuttavia, ha sollevato notevoli preoccupazioni tra gli esperti del settore. Yoshua Bengio, professore di informatica all'Università di Montreal e figura di spicco nella ricerca sull'IA, ha lanciato un avvertimento sui potenziali pericoli di questi nuovi modelli. Se OpenAI effettivamente superasse un livello di rischio medio per le armi CBRN (chimiche, biologiche, radiologiche e nucleari), ha affermato, ciò non farebbe altro che rafforzare l'importanza e l'urgenza di adottare una regolamentazione pubblica.

Oltre a lui, trentotto dipendenti, attuali ed ex, di OpenAI, Meta, Google DeepMind, Anthropic e xAI hanno firmato una lettera aperta a sostegno del disegno di legge del Senato della California 1047 (SB 1047). Quest'ultimo, approvato da entrambe le camere della legislatura statale, riterrebbe gli sviluppatori di IA direttamente responsabili degli eventuali danni causati dai loro modelli qualora non fossero state adottate misure di sicurezza adeguate. La lettera dei dipendenti mette in guardia, difatti, dai gravi rischi che i modelli di IA più potenti potrebbero presto comportare come un accesso ampliato alle armi biologiche, attacchi informatici alle infrastrutture critiche, ecc.

In conclusione, considerata l'attualità di tale potenziale minaccia non si può che avallare la tesi dello sviluppo ulteriore di una IA antropocentrica possibile solo mediante l'implementazione di stringenti misure di salvaguardia da parte delle principali Big Tech.