Tecnologia - *Yoshua Bengio lancia una nuova sfida: creare un'intelligenza artificiale onesta*

Roma - 04 giu 2025 (Prima Notizia 24) **"È teoricamente possibile immaginare macchine prive di un sé, di un obiettivo, che siano solo macchine di pura conoscenza come uno scienziato che sa tante cose".**

Yoshua Bengio, professore all'Università di Montreal, vincitore del premio Turing, considerato come uno dei padri dell'Intelligenza Artificiale, lancia una nuova sfida: creare un'Ia che sia onesta. Per questo, ha creato un'organizzazione chiamata "Law Zero", il cui obiettivo è quello di sviluppare una forma 'onesta' di intelligenza artificiale, che possa identificare i sistemi che tentano di ingannare o danneggiare le persone. Da anni, Bengio mette in guardia sull'uso di questa tecnologia. "Attualmente, l'IA viene sviluppata per massimizzare i profitti - dice Bengio - Vogliamo costruire IA che siano oneste e non ingannevoli. È teoricamente possibile immaginare macchine prive di un sé, di un obiettivo, che siano solo macchine di pura conoscenza come uno scienziato che sa tante cose". Con un finanziamento iniziale di quasi 30 milioni di dollari e un team di una dozzina di ricercatori, l'organizzazione intende creare un nuovo sistema chiamato Scientist AI, una specie di "guardiano" che controllerà e impedirà comportamenti pericolosi da parte di agenti di IA autonomi. Rispetto agli strumenti attualmente disponibili, il sistema di Bengio non darà risposte definitive, ma fornirà delle probabilità sulla correttezza di una risposta. Questo progetto nasce mentre i modelli linguistici di grandi dimensioni di OpenAI, Google e Anthropic vengono attuati nell'economia digitale, pur se con problemi. Recentemente, l'azienda di Ia Anthropic ha dichiarato che urante i test di sicurezza il suo ultimo modello di IA ha tentato di ricattare un ingegnere, per evitare di essere sostituito da un altro sistema.

(Prima Notizia 24) Mercoledì 04 Giugno 2025