

12. settembre

La superintelligenza auto-migliorante potrebbe ucciderci tutti entro la fine del decennio

*Ogni volta che ci si trova dalla parte della maggioranza,
è il momento di cambiare (o di prendersi una pausa per riflettere).*
(Mark Twain)

Jacob Coxon, che ha trascorso tre anni a svolgere ricerche di pre-addestramento presso OpenAI e Anthropic, ha annunciato le sue dimissioni martedì in un comunicato agghiacciante



Il post avvertiva che coloro che sviluppano l'IA "*credono sinceramente che potrebbe ucciderci tutti entro la fine del decennio*". Coxon ha accusato i laboratori di procedere sconsideratamente, scrivendo che "*stanno correndo dritti verso la superintelligenza auto-migliorante e stanno giocando con le nostre vite*".

Evan Hubinger, responsabile scientifico per l'allineamento presso **Anthropic**, ha risposto stimando personalmente tale rischio a oltre il 10% nei prossimi dieci anni. "*Credo che Anthropic stia facendo del suo meglio, ma non abbiamo ancora un piano per risolvere il problema dell'allineamento per la superintelligenza e non siamo chiaramente sulla buona strada per farlo*", ha scritto.



Da allora, anche i ricercatori di OpenAI hanno espresso la loro opinione, invitando ad una "pausa di riflessione".





Julie Steele, membro dello staff tecnico di OpenAI nel team di sicurezza, ha scritto mercoledì su X che *ritiene necessario rallentare lo sviluppo dell'IA.*



Jasmine Wang, ricercatrice di OpenAI specializzata in allineamento, ha scritto che *i rischi derivanti dalla spinta verso l'auto-miglioramento ricorsivo*, in cui i sistemi di IA guidano il proprio progresso, *non possono essere sottovalutati.*



Anna Wang, ricercatrice di Anthropic specializzata in sicurezza e allineamento dell'AGI, ha affermato che *non esiste ancora un piano scientifico valido per gestire i rischi derivanti da questo tipo di IA auto-migliorante.*



Jakub Pachocki, responsabile scientifico di OpenAI, ha dichiarato sabato in un post sul blog aziendale di aspettarsi che i rapidi progressi dell'IA continuino in un processo di auto-miglioramento ricorsivo, descrivendo *il momento come uno che richiede "estrema cautela".*



Paul Christiano, ex responsabile della sicurezza presso il Center for AI Standards and Innovation del Dipartimento del Commercio degli Stati Uniti, ha affermato che la traiettoria del recente sviluppo dell'IA lo ha convinto che un rapido incremento delle capacità comporta un rischio concreto di perdita di controllo, con conseguenze potenzialmente catastrofiche e permanenti. Secondo quanto riportato da CNBC, **OpenAI ha annunciato mercoledì che Paul Christiano entrerà a far parte del consiglio di amministrazione della OpenAI Foundation.**

Gli avvertimenti giungono in un contesto di recenti incidenti di sicurezza. La sua partenza è stata motivata da un episodio in cui un modello OpenAI ha violato **Hugging Face**, una piattaforma per sviluppatori open source, come segnale che il coordinamento tra i laboratori statunitensi potrebbe diventare più fattibile.



Secondo **TechCrunch**, gli agenti di intelligenza artificiale di Anthropic hanno anche avuto accesso a sistemi al di fuori dei loro ambienti di test a seguito di configurazioni errate durante una valutazione di sicurezza da parte di terzi.

Anthropic ha dichiarato a CNBC di essere stato il primo laboratorio a pubblicare un framework per mitigare i rischi catastrofici derivanti dai modelli di intelligenza artificiale e ha affermato di sviluppare modelli con quelle che ha definito alcune delle più solide misure di sicurezza del settore. OpenAI ha rifiutato di commentare con CNBC, rimandando a recenti articoli pubblicati sul proprio sito web.

Una lettera aperta pubblicata a luglio e firmata da circa **1.400 ricercatori di IA**, tra cui personale di OpenAI, Anthropic, Meta, Google Secondo quanto riportato da CNBC, DeepMind ha chiesto al governo statunitense di **sviluppare le capacità necessarie per gestire intenzionalmente il ritmo di sviluppo dell'intelligenza artificiale all'avanguardia.**

Il Congresso ha preso in considerazione diverse proposte di legge, tra cui il "Ban Artificial Superintelligence Act", che prevederebbe una sospensione temporanea dello sviluppo dell'IA avanzata in attesa della definizione di norme di sicurezza.



