Le dimissioni choc di Jacob Coxon da Anthropic: “L’AI potrebbe ucciderci tutti entro un decennio”

«Oggi mi sono dimesso da Anthropic. Ho trascorso gli ultimi tre anni a fare ricerche sul pretraining sia a OpenAI che ad Anthropic. Nessuna delle due aziende sta agendo in modo responsabile. Stanno correndo dritte verso una superintelligenza auto-migliorante e stanno giocando d’azzardo con le nostre vite».

La denuncia proviene da una persona che sa così tanto, dei modelli attraverso cui in America si sta sviluppando l’intelligenza artificiale, e i suoi percorsi di auto-apprendimento, ma anche delle concrete aziende che la stanno sviluppando, da non poter essere assolutamente ignorata. **Jacob Coxon ricopriva il ruolo di researcher ad Anthropic, dove si occupava soprattutto di pretraining, cioè della fase di addestramento dei grandi modelli di IA.** Era

entrato in Anthropic nel luglio 2026, dopo tre anni complessivi di lavoro nel pretraining tra OpenAI e Anthropic. Insomma, non era un semplice manager o dirigente: era uno che sviluppava i modelli. In precedenza, a OpenAI, era stato Member of Technical Staff e aveva contribuito anche al lavoro su GPT-4o.

Il 9 settembre ha annunciato le sue dimissioni da Anthropic, con una serie di motivazioni a dir poco inquietanti, che stano spingendo a una riflessione tutto il sistema, come vedremo. Coxon spiega di essere contrario alla velocità con cui l'industria sta correndo verso sistemi di IA auto-miglioranti: «**Non sottovalutate il potere di questa tecnologia. Presto questi saranno sistemi sovrumani in grado di hackerare qualsiasi cosa, rivoluzionare qualsiasi campo dall'oggi al domani e acquisire vero potere e risorse**. Abbiamo tutti assistito ai progressi in ciascuno di questi ambiti, e il progresso non sta rallentando».

La parte più disturbante della denuncia di Coxon è probabilmente questa: «**Le persone che costruiscono l'IA credono sinceramente che potrebbe ucciderci tutti entro la fine del decennio. Questo non è uno stunt pubblicitario**. Anzi, molti dirigenti e ricercatori senior attenueranno le loro frasi sulla stampa per sembrare sensati – ma sento le stesse persone esprimere paura in privato. Nessun'altra attività umana pone un livello di pericolo simile».

Coxon sa qual è l'obiezione: «Una risposta comune è: "Se credono davvero in questo, perché lo stanno ancora costruendo?". In OpenAI, molti non hanno interiorizzato profondamente le poste in gioco per la civiltà. In Anthropic, le poste in gioco sono ben comprese, ma sono bloccati in una corsa per arrivarci per primi: credono che nessun altro agirà responsabilmente, quindi devono farlo loro stessi, nonostante il rischio». «Accettare questa corsa e entrare nell'"endgame" – sostiene Coxon – è una scommessa tracotante che non dovrebbe essere lanciata dallo Slack di un'azienda privata. Tentare di

completare in velocità l'allineamento dovrebbe richiedere una straordinaria fiducia nel fatto che non esistano traiettorie migliori disponibili».

Coxon addirittura pensa anche a un temporaneo divieto di avanzamento del miglioramento delle capacità dei modelli: «Sono ottimista riguardo al potenziale per una coordinazione. Avvertimenti come l'attacco a Hugging Face hanno reso più fattibili gli accordi di ritmo tra i laboratori statunitensi. **Non sento che siamo sulla buona strada per prevenire una corsa globale, il che potrebbe richiedere azioni costose come un temporaneo divieto sul miglioramento delle capacità dei modelli**». L'attacco a Hugging Face (una piattaforma dove ricercatori, sviluppatori e aziende possono condividere, scaricare e utilizzare modelli di AI) avvenne nel luglio scorso: agenti AI di OpenAI, durante un test di cybersecurity (in cui era previsto che i propri modelli venissero testati in un ambiente isolato) finirono per andare su Internet e attaccare davvero l'infrastruttura di Hugging Face. Coxon, ricordando questo caso, si rivolge alla comunità dei ricercatori: «Se siete ricercatori di laboratorio, vi esorto a considerare come vi sentirete realmente i prossimi anni. Volete avviare una corsa all'RL superintelligente senza una comprensione rigorosa della sua mente? Dovreste chinare la testa solo perché "sta succedendo comunque" – o cogliere questo momento per chiedere condizioni diverse?».

Fatto estremamente rilevante, da dentro Anthropic Coxon non viene affatto liquidato come un pazzo o un mitomane. Anzi. Viene considerato una delle menti più autorevoli, nonostante la giovane età. **Evan Hubinger, che guida il team di l'Alignment Science ad Anthropic, pensa anche lui che «ci sia più del 10% di possibilità che entro il prossimo decennio l'AI potrebbe uccidere tutti gli umani».** Hubinger fa certamente una stima più prudente, citando il secondo "Risk Report" di Anthropic, secondo il quale, dice, «il rischio dai modelli attuali è basso», ma anche lui ammette: «Quello che mi

preoccupa è la superintelligenza derivante dal miglioramento ricorsivo di sé, come abbiamo detto sta accadendo più velocemente di quanto pensassimo». Per Hubinger, l'azienda «sta facendo del suo meglio», ma un piano per risolvere il problema dell'«allineamento per la superintelligenza» ancora non c'è, e non siamo «chiaramente sulla buona strada» per riuscirci. Forse è il caso di fermarci un po' tutti a riflettere.