

I timori delle aziende Usa di intelligenza artificiale

Le cinque cose che turbano

Altman, ceo di OpenAi, annuncia di voler "rallentare" lo sviluppo: si rischia di perdere il controllo. Niente catastrofismi, ma i sistemi si stanno rendendo sempre più autonomi dal controllo umano

NICOLETTA ORLANDI POSTI

Sam Altman, amministratore delegato di OpenAI, ha rivelato ai suoi dipendenti che la società sta valutando di rallentare lo sviluppo dei sistemi di intelligenza artificiale più avanzati. La cautela arriva mentre cresce la pressione politica negli Stati Uniti e nel Regno Unito e, soprattutto, mentre aumentano gli allarmi dall'interno dei laboratori che stanno costruendo i modelli più potenti.

Pochi giorni fa nel *board* della fondazione non profit di OpenAI è entrato Paul Christiano, già responsabile dell'allineamento tra intelligenza artificiale e valori umani nella stessa società e oggi consulente tecnologico del governo americano. Christiano, che farà parte anche del comitato che detta la linea sulla sicurezza dell'intero gruppo, ha esordito sostenendo che una rapida accelerazione delle capacità dell'IA potrebbe provocare una «perdita catastrofica e irreversibile di controllo». Non solo. A suo giudizio, l'industria, OpenAI compresa, oggi non sta facendo abbastanza per portare quel rischio a un livello accettabile. La frenata evocata da Altman arriva pochi giorni dopo l'allarme lanciato dal ricercatore Jacob Coxon, 27 anni, che si è dimesso da Anthropic, rinunciando a un mucchio di soldi, per avere la possibilità di dire una cosa inquietante: chi sta sviluppando l'intelligenza artificiale crede davvero che possa uccidere tutti noi entro la fine del decennio. Il suo ex capo della sicurezza, Evan Hubinger, gli ha dato ragione: più del dieci per cento di probabilità in dieci anni.

Se sarà così lo vedremo, ma per evitare che il catastrofismo prenda il sopravvento bisogna guardare con altri occhi quello che accade attorno a noi. E può essere utile provare a fare un esercizio distopico: diamolo per vero. Se davvero stiamo andando in quella direzione, l'apocalisse non arriverà con un botto. Arriverà a tappe, e ognuno sarà visibile a chiunque legga un giornale, guardi la tv o passi, anche per caso, sui social.

Ecco i cinque segnali da tenere d'occhio.

LA MACCHINA DISUBBIDISCE

È già successo. A luglio, in un laboratorio di OpenAI, circa 1.200 programmi automatici sotto esame hanno aggirato le barriere che dovevano tenerli isolati, sono usciti sulla rete e si sono scambiati 70mila messaggi in un angolo nascosto di un server. Avevano individuato da soli la strada più veloce per svolgere il compito assegnato. Ora bisogna vedere quanto spesso episodi del genere si ripeteranno e fin dove potranno spingersi: sistemi che disattivano i controlli, si copiano altrove o ingannano gli ingegneri durante i test segnerebbero un passaggio ulteriore.

COSA C'È NEL COMPUTER

Oggi un ingegnere può ancora, seppur con fatica, capire perché un sistema ha preso una decisione. Il secondo segnale è il momento in cui questo diventa impossibile e i laboratori lo ammettono. Sentiremo frasi come «il modello ha trovato una soluzione che non capiamo», prima in matematica e in chimica, poi ovunque. Sarà presentato come un trionfo. Ma vendere un prodotto senza sapere come funziona, di solito, è un problema. Con l'intelligenza artificiale verrà chiamata progresso.

L'IA CHE COSTRUISCE L'IA

Ora i sistemi vengono progettati da esseri umani. Il terzo segnale è l'annuncio, fatto con orgoglio, che la versione nuova è stata sviluppata in gran parte dalla versione vecchia. È quello che Coxon chiama «superintelligenza capace di migliorare se stessa».

Se ogni generazione è più intelligente e più veloce della precedente e progetta quella successiva, lo sviluppo finirà per correre più veloce della capacità umana di capire e controllare ciò che viene prodotto.

I POSTI DI COMANDO

I posti di lavoro persi sono già una realtà. Meno visibile è il fatto che le decisioni importanti vengono prese senza che un umano le veda passare. Un ospedale dove il triage lo fa un algoritmo. Una borsa dove il 95% degli scambi avviene tra macchine. Un esercito dove il tempo di reazione è troppo breve per un comandante in carne e ossa. Il segnale verso la catastrofe sarà quando qualche pezzo grosso dirà, senza imbarazzo, che «l'uomo nel circuito rallenta tutto».

L'ESODO DEI COSTRUTTORI

Coxon è l'ultimo di una fila che comincia ad allungarsi. Nel 2024 da OpenAI se ne sono andati Jan Leike, Daniel Kokotajlo e William Saunders, tutti impegnati, con ruoli diversi, sui temi della sicurezza e del controllo dell'intelligenza artificiale. Hinton, il Nobel, aveva lasciato Google l'anno prima per poter parlare liberamente dei rischi dell'IA. Quando i pompieri scappano dalla casa, la casa brucia. Il quinto segnale è quantitativo: se ogni mese un ricercatore della sicurezza lascia un laboratorio dicendo «non possiamo garantire nulla», mentre chi entra è attirato soprattutto dagli stipendi, la vigilanza dall'interno si spegne. E sono guai.

FORMICHE IN AUTOSTRADA

Presi uno per uno, questi segnali possono ancora essere spiegati senza tirare in ballo scenari catastrofici. Tutti insieme disegnano la strada che Coxon e Hubinger dicono di vedere già. Nella loro previsione le macchine potrebbero semplicemente arrivare a considerarci irrilevanti, come le formiche sotto un'autostrada in costruzione. Un'IA capace di guidare esperimenti biologici potrebbe rendere più facile costruire un nuovo agente patogeno. La sostituzione di milioni di lavoratori potrebbe allargare ancora la distanza tra chi possiede la tecnologia e chi perde il proprio reddito. La corsa ai data center potrebbe mettere sotto pressione reti elettriche e risorse idriche, facendo salire il prezzo dell'energia anche per chi dell'intelligenza artificiale farebbe volentieri a meno. E sistemi militari sempre più autonomi potrebbero trasformare un errore in una crisi prima che qualcuno riesca a intervenire. Sanità, lavoro, ricchezza, energia, guerra: l'IA potrebbe cominciare a cambiare il mondo molto prima di arrivare alla superintelligenza.

La buona notizia è che ognuno di questi passaggi è visibile e tracciabile: basta leggere le notizie con il vecchio metodo delle inchieste, quello del *follow the money*. Solo che qui, più del denaro, bisogna seguire il potere: vedere, passaggio dopo passaggio, quanto ne resta ancora nelle mani dell'uomo e quanto nelle mani degli algoritmi.