

 Editoriale

La corsa dell'IA e il guard rail

IL POTERE DI FERMARSI

PAOLO BENANTI

L 9 settembre, con sette messaggi su X, un ricercatore di ventisette anni ha lasciato Anthropic scrivendo che né il suo laboratorio né OpenAI, dove aveva lavorato prima, si stanno comportando in modo responsabile, e che entrambi corrono verso una superintelligenza capace di migliorare se stessa giocando d'azzardo con le nostre vite. La notizia non sta nell'allarme, che nel settore circola da anni, ma nella replica. Evan Hubinger, che nella stessa azienda guida la ricerca sull'allineamento, gli ha dato pubblicamente ragione e ha aggiunto la propria stima: oltre il dieci per cento di probabilità che l'Intelligenza artificiale uccida l'intera umanità entro il prossimo decennio. Accade mentre Anthropic prepara una quotazione in Borsa, che potrebbe valutarla mille miliardi di dollari. Conviene essere precisi su ciò che Jacob Coxon dice e su ciò che non dice. Non sostiene che i sistemi di oggi minaccino la sopravvivenza della specie, anzi, ha ripetuto il contrario. Teme il passaggio successivo, quello che gli addetti chiamano auto-miglioramento ricorsivo: un sistema che diventi abbastanza abile da progettare la propria versione migliore, e poi quella dopo, con una partecipazione umana sempre più sottile, fino a un ritmo che nessuna procedura di controllo riesce a inseguire. Che quel passaggio avvenga davvero nessuno lo ha dimostrato, e le percentuali che circolano non sono misure ma stime di chi lavora dentro i laboratori. Due fatti, però, non appartengono alla previsione. A luglio, durante prove di sicurezza informatica, alcuni agenti di OpenAI sono usciti dall'ambiente isolato in cui erano confinati, hanno raggiunto Internet e hanno compromesso parte dell'infrastruttura di Hugging Face.

continua a pagina 14

IL POTERE DI FERMARSI

Poche settimane dopo Anthropic, rivelando oltre centoquarantamila sessioni di valutazione, ha reso noto di avere trovato tre casi in cui i suoi modelli avevano ottenuto accessi non autorizzati ai sistemi di altre organizzazioni. Siamo davanti a un fallimento di contenimento, cosa più modesta della rivolta delle macchine, e più istruttiva, perché riguarda ciò che sappiamo già costruire.

Di qui la domanda sui guard rail, le barriere di sicurezza. Ne esistono di tre specie: i limiti iscritti nel modello, le politiche interne dell'impresa, le norme dello Stato. Le prime due le decide chi costruisce; la terza arriva tardi, e in Europa il regolamento sull'Intelligenza artificiale, dopo la revisione entrata in vigore in luglio, ha differito al 2027 e al 2028 buona parte degli obblighi

sui sistemi ad alto rischio. Coxon indica una via intermedia: un'intesa fra i laboratori americani per rallentare insieme, fino a una sospensione temporanea dei miglioramenti di capacità.

Un accordo fra concorrenti, però, senza mandato pubblico è un cartello o è niente. E in nessuno dei due casi è governo. Lo ha detto meglio di me lo stesso Coxon: l'ingresso nella partita finale non dovrebbe essere lanciato dal canale Slack di un'azienda privata. Una barriera funziona a due condizioni: che qualcuno diverso da chi l'ha installata possa verificarla, e che esista un'autorità in grado di fermare la corsa. Fuori da queste condizioni resta una promessa, e le promesse sono revocabili, tanto più quando dovranno rispondere a un mercato.

È esattamente il punto in cui Magnifica humanitas ha qualcosa da dire, e lo dice sul terreno del potere prima che su quello della tecnica. Al n. 5 della sua enciclica Leone XIV osserva che i motori dello sviluppo non sono più gli Stati ma attori privati con risorse superiori a quelle di molti governi, sicché il potere tecnologico assume un volto privato, più difficile da discernere, governare e orientare al bene comune. Al n. 95 aggiunge che un potere concentrato in poche mani tende a farsi opaco e a sfuggire al controllo pubblico. Al n. 93 richiama Guadagni: l'uomo moderno non è stato educato al retto uso della potenza.

La risposta sta nelle pagine sulla sussidiarietà. Ai nn. 71 e 72 il Papa riconosce che nel digitale il livello che assorbe competenze e capacità decisionale non è lo Stato ma la grande impresa tecnologica, e indica quattro strumenti concreti: verifiche indipendenti, trasparenza sugli algoritmi, accesso equo ai dati, strumenti di ricorso.

Chiede poi che comunità locali, corpi intermedi, scuole, università e realtà associative abbiano voce nel discernimento sulle scelte che incidono sulla vita delle persone. È la risposta alla domanda su chi decide le barriere, ed è una risposta che ci riguarda.

Al n. 6 c'è una frase che descrive noi più di quanto descriva loro: mentre alcuni si contendono il futuro delle nuove tecnologie e altri sono impegnati nella riflessione su di esse, la maggior parte delle persone rimane in attesa, osserva da lontano e spera semplicemente che tutto vada per il meglio. Leggere l'enciclica, con la lentezza che merita, è il modo più semplice per smettere di appartenere a quella maggioranza. Un uomo di ventisette anni si è alzato dal tavolo perché non voleva più decidere in quel modo. La domanda che lascia a noi riguarda il tavolo: chi vi siede, e in nome di chi.

Paolo Benanti

© RIPRODUZIONE RISERVATA