Sabato 12 Settembre 2026Corriere della Sera

I TRE MODI CON CUI L'AI PUÒ MINACCIARE L'UMANITÀ

di Massimo Gaggi

Valori, autotutela, perdita di controllo: i rischi estremi

Danni irreparabili all'umanità entro il 2030? Possibile estinzione del genere umano tra un decennio? I timori che hanno spinto Jacob Coxon di Anthropic e altri scienziati che costruiscono e addestrano nella Silicon Valley i modelli più potenti di intelligenza artificiale ad abbandonare i loro incarichi per non diventare i creatori di tecnologie incontrollabili e potenzialmente micidiali, fanno discutere, spaventano.

Rischi reali o esagerazione di «doomers», gente che si lascia andare a previsioni apocalittiche? E come pensano che possano materializzarsi questi rischi esistenziali?

Bé, il «doomerism» non nasce oggi. Si discute da anni della possibilità che errori di programmazione di macchine sempre più potenti ma prive di senso etico e di una scala di valori umani portino alla soppressione di molta di noi, o addirittura di tutta l'umanità: «Un danno collaterale nello svolgimento della missione ricevuta», secondo la definizione dello scienziato «pentito» di Berkeley, Stuart Russell. E già tre anni fa il premio Nobel Geoffrey Hiton, capo degli scienziati di Google, si dimise dall'azienda proprio per essere libero di denunciare i rischi estremi che vedeva insiti in uno sviluppo dell'AI talmente veloce da eccedere le capacità umane di comprensione. È stato lui uno dei primi a provare a quantificare il rischio di estinzione dell'umanità: 10 per cento.

Cosa immagina chi parla di rischi esistenziali? Il capostipite è il filosofo di Oxford Nick Bostrom che in Superintelligence, suo libro del 2014, ipotizzò uno scenario paradossale, ma da allora divenuto paradigmatico: il mondo accidentalmente distrutto dall'ordine di massimizzare la produzione di graffette, le clip di metallo per i fogli, dato con superficialità a un'intelligenza artificiale più potente dell'uomo. Potente, ma, in quanto priva di senso morale, di empatia, estranea alla difesa dell'umanità e dei suoi valori, la macchina prima assorbe tutta la produzione siderurgica mondiale, poi, scoprendo che negli atomi dei quali siamo fatti ci sono gli elementi coi quali si può produrre ferro, e quindi, clips, ci uccide tutti e trasforma il mondo in una fabbrica di graffette.

Fantasie fantascientifiche di un filosofo allora lette e diffuse, più con compiacimento che con preoccupazione, dai guru delle aziende di big tech. Dodici anni dopo, però, davanti ad AI sempre più potenti, i cui meccanismi di funzionamento sfuggono anche ai loro creatori che non riescono più a controllarle, quelle fantasie cominciano a prendere la forma di minacce da non trascurare.

Quali potrebbero essere le insidie per l'umanità? Gli esperti vedono almeno tre categorie di rischi:

1) Il disallineamento: se non si ha la capacità di accompagnare un ordine dato a una macchina con il suo allineamento ai valori umani e ai nostri obiettivi di fondo, la macchina può pensare che la soluzione del problema sia nella nostra soppressione. Esempio classico: un'AI alla quale viene ordinato di migliorare il clima può concludere che il modo migliore di farlo è quello di eliminare tutti gli inquinatori, cioè noi.

2) Autotutela. Stiamo scoprendo che le macchine, una volta ricevuto un ordine, si preoccupano di eliminare i fattori che potrebbero bloccarle. Potrebbero, quindi, vedere un ostacolo anche negli umani.

3) Qui entra più direttamente il fattore umano col rischio di affidare alla gestione dell'AI le tecnologie più pericolose e complesse, perdendone poi il controllo: la gestione degli arsenali nucleari, la capacità di sperimentare a velocità siderale nuove combinazioni biochimiche che possono diventare farmaci salvavita o agenti patogeni per armi biologiche micidiali, la gestione di infrastrutture essenziali (acqua, elettricità, trasporto aereo eccetera) che possono essere esposte a cybersabotaggi.

Quanto sono concreti questi rischi? Difficile dirlo, ancor più fare il gioco delle percentuali. Ormai sono tanti, però, gli scienziati che parlano di rischi reali, ammessi in parte anche dalle imprese. C'è, però, anche chi avverte che pensando a scenari apocalittici si perdono di vista rischi meno catastrofici ma immediati: l'uso dell'AI per produrre disinformazione, stravolgere la politica. E poi, oltre ai lavori che scompaiono, l'atrofia cognitiva di chi si affida sempre più a sistemi automatici che ci sostituiscono anche nel pensiero critico, nelle decisioni.

Prende forma, così, una nuova categoria di «doomers»: più che alla soppressione del genere umano per un singolo errore di programmazione delle macchine pensano al rischio di finire come la rana in una pentola con la temperatura che sale gradualmente: un sempre più diffuso affidarsi ad analisi e decisioni automatizzate, dalla gestione di campagne militari a quelle delle reti essenziali per la vita sociale: una catena di decisioni apparentemente ragionevoli che, però, porta a un collasso. Va risolto il problema dell'allineamento: rallentare (con accordi, per nulla facili, tra imprese e tra Stati) lo sviluppo dell'AI per dare tempo agli esperti in carne ed ossa di costruire le necessarie salvaguardie, ancorando solidamente le macchine alle esigenze dell'uomo, scongiurando i pericoli.