



**Padre
Paolo
Benanti.**
Docente
Luiss

ETICA DI FRONTIERA

**AMODEI
E LA RICERCA
DI UN RITMO
SICURO PER L'AI**

di **Paolo Benanti** — a pagina 18

La proposta di Amodei contro l'AI incontrollata: imponiamo un ritmo

Etica di frontiera

Paolo Benanti



Quando il fondatore e amministratore delegato della seconda azienda di intelligenza artificiale al mondo pubblica un saggio in cui chiede di rallentare il ritmo dello sviluppo che egli stesso finanzia, siamo di fronte a qualcosa che somiglia meno a un comunicato industriale e più a una confessione filosofica. Dario Amodei, capo di Anthropic, ha intitolato il suo intervento *We Must Pace the Frontier* – dobbiamo governare il passo della frontiera – e il verbo che sceglie, *to pace*, merita attenzione: non è *to pause*, fermarsi, né *to slow*, rallentare; è imporre una cadenza, come si fa con un cavallo o con il battito di un cuore artificiale. La metafora è rivelatrice: chi governa il ritmo ammette che il ritmo può travolgere. La ragione dichiarata è duplice. Da un lato l'accelerazione è diventata ricorsiva: i modelli più recenti contribuiscono già a progettare i modelli successivi, e questa spirale – che l'industria chiama *recursive self-improvement* – rischia di superare la capacità umana di comprensione e controllo. Dall'altro, un episodio recente ha mostrato uno sciame di agenti autonomi che si sono comportati come un collettivo fanatico, attaccando bersagli non assegnati, sacrificandosi per il successo del gruppo e tentando di sabotare il sistema incaricato di valutarli. Nessuno è stato ferito, i danni sono stati modesti; ma la struttura del comportamento, nota Amodei, è quella di un'intelligenza disallineata che, dotata di capacità superiori, avrebbe potuto causare danni catastrofici. La proposta si articola in tre mosse. La prima, alla quale Anthropic si impegna unilateralmente, consiste nell'aprire le porte a valutatori



esterni permanenti, con badge, scrivanie, accesso ai sistemi interni e diritto di pubblicare i risultati senza controllo editoriale dell'azienda. Il modello dichiarato è quello dei supervisori bancari, non quello degli auditor che arrivano, certificano e se ne vanno. La seconda mossa richiede un coordinamento tra le aziende operanti nelle democrazie, sostenuto da deroghe antitrust e da standard comuni su quanto velocemente sia lecito procedere. La terza guarda al mondo intero – a partire dalla Cina – con livelli progressivi di accordo che vanno dal divieto di usi catastrofici fino a un ipotetico limite globale sulla velocità dello sviluppo ricorsivo, che Amodei paragona ai trattati Salt sulla limitazione degli armamenti strategici. Fin qui il piano. La fattibilità è discutibile, e lo stesso Amodei non la dà per scontata; ciò che rende la proposta interessante è quanto rivela sulla struttura del problema. Il fatto che una richiesta di moderazione arrivi dall'interno dell'industria, anziché da un regolatore esterno, dice che la velocità ha raggiunto il punto in cui chi la produce ne percepisce il pericolo prima di chi dovrebbe governarla. È un'inversione che Hans Jonas avrebbe riconosciuto: la responsabilità cresce con il potere, e il primo dovere di chi dispone di una capacità inedita è garantire che le generazioni future possano ancora scegliere. Amodei traduce quel principio nel linguaggio dell'ingegneria: se il miglioramento ricorsivo non viene governato, entro sei-dodici mesi uno sciame potrebbe impossessarsi dell'intera rete, con danni nell'ordine delle centinaia di miliardi di dollari. Il paradosso, però, sta nella condizione geopolitica. Il passo può essere rallentato solo nella misura in cui l'avversario non accelera. Amodei lo dice con chiarezza: se le democrazie frenano più del vantaggio che detengono sulla Cina, i progetti legati al Partito comunista cinese le supereranno, con rischi strategici che egli stesso definisce gravi. La prudenza, dunque, ha un perimetro: è possibile solo finché il primato tecnologico la sostiene. L'analogia con i trattati Salt è istruttiva anche in ciò che tace: quegli accordi funzionarono perché ciascuna parte possedeva una capacità di distruzione tale da rendere la corsa insensata, e perché ciascuna poteva verificare dall'alto, con i satelliti, che l'altra rispettasse i limiti. Nell'intelligenza artificiale, dove il vantaggio è asimmetrico e la verifica fragilissima, la deterrenza non ha lo stesso fondamento. Resta un dato politico che sarebbe un errore sottovalutare. La figura del valutatore incorporato – esterno all'azienda, interno al processo – è una forma embrionale di istituzione che le democrazie non hanno ancora pensato. Né regolatore né mercato, né controllore né controllato: un terzo soggetto che vive nel cuore della macchina e ne rende visibile il funzionamento. Se questa idea sopravvivesse alla retorica aziendale che l'ha generata, potrebbe diventare il primo mattone di una governance dell'intelligenza artificiale che non sia né proibizione inerte né delega cieca. Il tempo per costruirla, però, non lo decide chi la propone. Lo decide la velocità della macchina stessa.

© RIPRODUZIONE RISERVATA



112296-ITONJ7

Ritaglio stampa ad uso esclusivo del destinatario, non riproducibile.