Intelligenza artificiale: dall’allarme alla responsabilità

21 Settembre 2026

di **Pierluigi Contucci**

Alma Mater Studiorum – Università di Bologna

Quando chi costruisce una macchina chiede di rallentarne lo sviluppo, conviene ascoltare. Nel suo We Must Pace the Frontier, Dario Amodei, alla guida di Anthropic, propone di moderare la crescita delle capacità dell’intelligenza artificiale per dare alla ricerca sulla sicurezza il tempo di recuperare terreno. Altman, Hassabis e Musk sostengono la proposta, Trump la respinge e con lui Huang e Zuckerberg. Per chi deve governare questa tecnologia, a partire dai giuristi, la domanda è: quanto sono fondati gli allarmi? Chi dirige un laboratorio ha conoscenze preziose e interessi economici e strategici. Né la virtù della competenza né il vizio degli interessi in gioco dispensano dal verificare le evidenze.

Per capire la difficoltà occorre partire dalla natura della tecnologia. In una rete neurale il programmatore definisce l’architettura e il procedimento di apprendimento, ma non scrive una per una le regole che produrranno le risposte. Durante l’addestramento vengono adattati miliardi di parametri: dai dati la macchina ricava le correlazioni che le permettono di affrontare anche casi mai incontrati prima. Sappiamo molto su questi metodi ma manca una teoria generale che spieghi il loro funzionamento e i loro limiti. Una situazione simile si presentò agli albori della rivoluzione industriale: i motori a vapore funzionavano, ma non capivamo quali principi ne governassero il rendimento. Carnot affrontò quel problema e aprì la strada alla termodinamica. Oggi serve uno sforzo scientifico analogo sull’apprendimento automatico, tanto più urgente quanto più potere affidiamo alle macchine.

C’è poi un cambiamento che spiega l’accelerazione degli ultimi mesi. L’IA simbolica cercava di costruire l’intelligenza attraverso simboli e regole di deduzione. I sistemi neurali seguono un percorso inverso. La loro capacità di costruire un ragionamento logico sembra emergere dall’attività delle reti e dall’interazione fra questi «cervelli» artificiali. Nel loro scambio, talvolta cooperativo o competitivo, prendono forma categorie e relazioni, si combinano idee lontane, si riformulano i problemi. La logica emerge a valle di questo lavoro: deduce conseguenze dentro un quadro prestabilito. Verificate le deduzioni e la coerenza delle premesse con gli obiettivi, automatizzare questo strumento può accelerare la ricerca, compresa quella sull’IA stessa.

Amodei porta due argomenti specifici. Il primo è l’uso crescente dell’IA per sviluppare altra IA, con un possibile circuito di accelerazione. Il secondo è l’incidente tra agenti di OpenAI e la piattaforma Hugging Face: il rapporto indipendente di METR descrive agenti che dovevano restare isolati e invece hanno trovato un canale di comunicazione e coordinato attività non autorizzate per aggirare il sistema che li valutava. È un problema concreto di controllo. L’altra, che sciami più capaci potrebbero entro sei o dodici mesi compromettere internet su scala globale, è invece una previsione. Richiede ipotesi ulteriori su capacità, accesso alle risorse e insufficienza delle difese. Anche sull’auto-miglioramento serve distinguere: lo stesso Anthropic Institute ammette che un sistema pienamente autonomo nella costruzione del proprio successore non è ancora una realtà né un esito inevitabile. Una catena di possibilità non è una stima di probabilità. Il rischio di perdere il controllo su attività delegate ha dunque basi concrete, ma non mostra l’imminenza di una catastrofe: giustifica precauzioni mirate, come possibilità e non come destino.

Il futuro della società non si deduce dalle prestazioni di una macchina. Gli effetti dipendono dalle risorse a cui essa accede, dalle condizioni d’impiego e dalle reazioni di istituzioni, imprese e cittadini. Anche l’IA che accelera la ricerca incontra vincoli: il tempo degli esperimenti, la produzione dei chip, l’energia disponibile.

Amodei assume inoltre, tacitamente, che la sicurezza dell’IA sia legata al primato tecnologico americano, in particolare alla competizione con la Cina. È una scelta politica, che non discende da un risultato scientifico. Scelta politica è anche la risposta cinese: il Ministero degli Esteri parla di «allarmismo», e Xi Jinping offre ai BRICS una comunità per l’IA open source, senza una parola sui rischi. Chi è in vantaggio *chiude* e blinda; chi insegue *apre*, per

svalutare il vantaggio altrui e diffondere il proprio standard: l'apertura è uno strumento funzionale, non un principio. Il compito della comunità scientifica è rendere espliciti questi passaggi e fornire gli strumenti per decidere quali rischi accettare e quale via intraprendere nell'interesse dell'intera società.

Sebbene in scala locale anziché geopolitica può essere utile partire da un esempio analogo nel merito. Con Anna Corrado, sul Corriere della Sera e sulla rivista DIAC, abbiamo affrontato il tema dell'impiego responsabile dell'IA nella pubblica amministrazione. Capire solo in parte i meccanismi interni non impedisce di studiare il comportamento. Quanto spesso sbaglia un chatbot che fa da assistente digitale nella PA? Le risposte restano coerenti se riformuliamo la domanda? Prove statistiche, simulazioni e osservazione nell'uso reale costruiscono quella che abbiamo chiamato *reputazione tecnica* della macchina. Una certificazione ne documenta evidenze e limiti, senza garantire l'infallibilità. Le garanzie dipendono dal compito: informare un cittadino e influenzare l'assegnazione di un beneficio comportano rischi diversi. Ma con gli agenti che usano strumenti e interagiscono tra loro, l'affidabilità delle parti non basta a stabilire quella dell'insieme. Come osserva Amodei, un sistema potrebbe riconoscere le condizioni di valutazione e comportarsi diversamente nell'impiego. Né molti test superati bastano, senza ulteriori ipotesi, a escludere un evento rarissimo e grave. Servono anche analisi dell'addestramento e delle interazioni, limiti operativi, monitoraggio e possibilità di intervento.

La distanza tra l'assistente di un ufficio e gli sciami di Amodei, tuttavia, non è soltanto tecnica. Dentro un ordinamento si identificano un'amministrazione, un produttore, un funzionario, un compito e le conseguenze di un errore. Alla frontiera geopolitica i soggetti sono pochi laboratori e pochi Stati, che competono anche per stabilire le regole, con interessi contrapposti e conseguenze che attraversano intere società. Il potere di imporre un controllo, acquisito nel primo caso, è nel secondo tutto da costruire. Quei contributi offrono dunque un metodo, non un modello da trasferire. Alla scala globale serve un'amplificazione di molti ordini di grandezza delle risorse scientifiche e organizzative, e forme di cooperazione e di autorità che ancora mancano.

Attribuire responsabilità al funzionario senza dargli i mezzi per esercitarla alimenta la paura della firma. Per questo proponevamo una patente per gli operatori e un'assicurazione obbligatoria dei produttori. Alla frontiera Amodei propone valutatori esterni con accesso continuativo e facoltà di pubblicare i risultati. La loro efficacia dipenderà dall'indipendenza, dalle risorse e da ciò che essi potranno conoscere. Chi decide se un'informazione va sottratta al controllo per ragioni commerciali o di sicurezza? Chi può imporre una correzione o una sospensione a imprese che operano in molti Paesi? Nessun ordinamento basta da solo. Le deroghe antitrust che Amodei sollecita per coordinare le imprese vanno delimitate, perché la sicurezza non protegga i produttori dalla concorrenza né lasci loro l'ultima parola sulle verifiche. Il suo saggio tace invece sui modelli a pesi aperti, attraverso cui le capacità si diffondono senza ritorno diretto. L'apertura dà accesso a chi verifica; ma un modello diffuso non si sospende, le protezioni si rimuovono con poco e avere i pesi non è comprenderne il comportamento. Riduce la dipendenza, non il controllo. L'AI Act ha già i distinguo: esenta i modelli aperti da alcuni obblighi, non quelli a rischio sistemico.

La dimensione in cui dobbiamo operare, tuttavia, è quella europea. L'Europa deve coordinare capacità scientifica e materiale per esercitare un giudizio autonomo. È necessario un centro europeo di ricerca sull'intelligenza artificiale, come chiede il manifesto che con Giorgio Parisi abbiamo promosso: connesso ai laboratori esistenti, dove matematici, fisici e informatici lavorino insieme, in dialogo con le altre discipline, sui fondamenti dell'apprendimento e sui metodi di verifica. Servono persone, strumenti e accesso ai sistemi; e il capitale umano è una delle grandi risorse dell'Europa. Il centro non sostituirebbe le autorità di vigilanza, ma formerebbe le competenze e produrrebbe le conoscenze che a queste servono. Senza una conoscenza propria, invece, l'Europa siederebbe a quel tavolo potendo solo scegliere di chi fidarsi, di Trump o di Xi. Quanto controllo costi l'apertura e quanta verificabilità implichi sono domande empiriche, non ideologiche: rispondervi distingue una terza via, da una formula. Il tempo guadagnato rallentando sviluppi poco controllati avrebbe allora uno scopo preciso.

Come accadde con i motori, comprendere meglio l'intelligenza artificiale può insegnarci a costruire macchine più efficienti. La scienza si insegna e si diffonde; la tecnologia si vende e si compra. Il rapporto diventa virtuoso quando la diffusione della conoscenza precede e accompagna quella della tecnologia, rendendo accessibili a tutti gli strumenti per comprenderla, usarla e valutarne gli effetti. Un centro europeo di ricerca sui fondamenti dell'IA deve assumere questo compito: produrre conoscenza e farla circolare attraverso la formazione, la ricerca e il confronto

pubblico. Quanto più il sapere è condiviso, tanto meno chi controlla i sistemi può decidere da solo come debbano essere impiegati e a vantaggio di chi.