



Capacità emergenti e opacità cognitive dell'IA: una sfida per la democrazia?

## Descrizione

I sistemi di intelligenza artificiale generativa (IA) hanno compiuto un salto qualitativo, per certi versi imprevedibile. Oltre a generare testi e immagini, i modelli più recenti mostrano capacità emergenti. Abilità non programmate esplicitamente che compaiono spontaneamente al crescere della scala computazionale. Ragionamento analogico, pianificazione multi-step, comprensione del contesto. Competenze considerate esclusivamente umane. Questo fenomeno pone interrogativi profondi e la questione della *governance* diventa urgente. L'Unione europea con l'*AI Act* ha tentato un primo inquadramento normativo ma la velocità dell'innovazione supera quella della legislazione.

Una prima considerazione verte sulla distinzione tra riproduzione intelligente e produzione dell'intelligenza. Ovvero tra IA ingegneristiche o riproduttive e IA cognitive o produttive. Secondo questa distinzione l'evidenza data da una potenza di calcolo di gran lunga superiore a quella umana significa un'alternativa certamente molto utile ma non si può negare che potrebbe rappresentare anche una prospettiva evolutiva, per certi versi inquietante. Cosa resterebbe della nostra identità incarnata? Stiamo assistendo alla progressiva migrazione del pensiero dall'umano all'astrazione del calcolo? Un pensiero incorporeo che, abbandonando il limite, si fonda nel computazionale? Siamo di fronte a una nuova forma di soggettività cognitiva o ci riflettiamo in uno specchio algoritmico scambiando l'eco per una voce e il riflesso per un volto?

## Capacità emergenti

Analizziamo il fenomeno delle [capacità emergenti](#). In pratica, a mano a mano che aumentano i parametri e i dati di addestramento, questi sistemi raggiungono una complessità tale da far emergere funzionalità inaspettate, non programmate esplicitamente dagli sviluppatori. Possiamo dire che assistiamo a un salto di stato qualitativo, in cui il sistema architetta in modo indipendente strategie risolutive mai tracciate prima, perfezionando spontaneamente la propria abilità nel generare risposte che sorprendono per rigore logico e complessità strutturale.

Giunti a questo punto, potrebbe essere lecito ritenere che l'emergere di tali capacità cognitive può essere considerato un punto di convergenza tra l'architettura dei modelli linguistici artificiali e il pensiero umano? Questa inaspettata evoluzione dell'IA rappresenterebbe il segnale che si stanno ricalcando i processi cognitivi profondi della mente umana?

Sul tema la posizione dei ricercatori è contrastante. Per alcuni stiamo assistendo a una svolta decisiva mentre, per altri, permangono dubbi in merito all'interpretazione del fenomeno. Concorde è la considerazione che la capacità emergente si rileva nei modelli più grandi mentre è assente nei modelli più piccoli. Vale a dire che non si tratta di un semplice affinamento di *skills* preesistenti, bensì di qualcosa qualitativamente nuovo. Il modello è imparato a farlo solo superando una soglia critica di scala: numero di parametri, qualità/quantità dei dati e degli algoritmi di *training*.

## Opacità cognitive

Ebbene, come descritte, a queste capacità emergenti corrispondono anche opacità cognitive. Significa che per quanto si possa perfino avere conoscenza di tutti i dati interni al sistema, non siamo in grado di ricostruire con chiarezza perché una certa decisione è stata presa. La motivazione è riconducibile ad una varietà di fattori. In particolare, l'ostacolo non è solo il vuoto informativo ma l'eccesso di complessità. Ci troviamo davanti a sistemi di IA dove con innumerevoli variabili e relazioni non lineari si producono effetti che insieme non riconducibili a una semplice somma delle parti. Una classica esemplificazione, a tal proposito, è riconducibile al volo di uno stormo di uccelli. Ogni singolo uccello segue regole semplici di volo ma il movimento collettivo dello stormo non può essere spiegato sommando i comportamenti dei singoli uccelli.

Significa che le tradizionali spiegazioni non sono più sufficienti. Se non si è in grado di dare una risposta alla classica domanda sul perché un sistema di IA ha assunto una determinata decisione (opacità cognitive), la questione si sposta sul piano funzionale. La decisione assunta dal sistema di IA è utile? Funziona? Insomma, si passa dal piano della esplicabilità a quello del mero funzionalismo. Proprio perché il sistema di IA ha operato all'interno di spazi decisionali inconoscibili, si sancisce il passaggio dalla cognizione effettiva alla pura osservazione passiva del risultato e dei suoi effetti.

Il cuore del problema, pertanto, consiste nel combinato disposto tra capacità emergenti e opacità cognitive, da cui la validazione funzionale con il seguente assunto: l'efficacia di un risultato conta più della comprensione del percorso che lo ha generato. Non è necessario chiedersi il perché, in quanto ci è tecnicamente impossibile saperlo, ma solo che l'output funzioni.

## Accettare risultati senza comprenderne il processo decisionale?

Sebbene questo approccio funzionalista consenta di gestire complessità altrimenti inaffrontabili, porta con sé rischi strategici ed etici profondi. Non è in grado di decifrare i criteri sottostanti, il ruolo dell'umano muta da decisore consapevole a mero fruitore che approva output nati in spazi che gli sono preclusi.

Lo slittamento, dalla comprensione alla passiva accettazione, è la frattura epistemica più insidiosa dell'IA in quanto le capacità emergenti e le opacità cognitive non viaggiano separati. Si alimentano a vicenda. E mettere in discussione un sistema, in grado di evidenziare capacità non

previste ma che *non funzionano*, si rischierebbe di apparire ostili al progresso, nostalgici di un'inefficienza ormai superata.

A fronte del funzionalismo della tecnica si pone il diritto a comprendere una decisione che ci riguarda. Diritto fondativo in qualsiasi sistema che si voglia chiamare democratico ma che le *opacità cognitive* vanificano.

La risposta più diffusa a questo problema sarebbe *Explainable Artificial Intelligence (XAI)* che consiste nello sviluppo di tecniche che possono rendere interpretabili le decisioni assunte dall'IA. È un orientamento certamente necessario ma non sufficiente. Le spiegazioni prodotte dai sistemi di XAI sono spesso approssimazioni *post-hoc*, non descrivono il processo reale di elaborazione ma costruiscono una narrazione plausibile a partire dal risultato. In un certo senso possiamo dire che rappresenta una *opacità di secondo livello*. Da un lato non sappiamo come il sistema ha sviluppato un determinato *output* ma abbiamo un risultato e una storia che sembrano ragionevoli. Il rischio è che quella storia diventi il surrogato della comprensione e che la *governance* si accontenti di esso.

La questione richiede un qualcosa di più profondo. Possiamo dire una nuova cultura della responsabilità digitale. Capace di tenere insieme ciò che l'IA sa fare e ciò che la società deve poter comprendere e governare. Nel neoumanesimo digitale questa è una sfida per la democrazia.

Crediti foto: Gertruda Valaseviciute su Unsplash

#### **Data di creazione**

22 Maggio 2026

#### **Autore**

lucio\_romano