Claude è un pericolo?

Quando l'IA comincia a costruire se stessa...

L'articolo di Anthropic; "When AI builds itself" ruota attorno a una domanda semplice solo in apparenza: che cosa succede quando l'intelligenza artificiale smette di essere uno strumento nelle mani degli esseri umani e diventa parte del processo che produce nuova IA? Non siamo ancora davanti a macchine pienamente autonome, ma a un cambiamento visibile: i modelli scrivono codice, eseguono esperimenti, correggono errori e partecipano allo sviluppo dei sistemi successivi.

Il concetto centrale è quello di miglioramento ricorsivo dell'IA: la possibilità che un sistema artificiale contribuisca alla progettazione e al perfezionamento dei modelli successivi. Anthropic chiarisce che lo scenario non è ancora realtà e non è inevitabile. Il punto dell'articolo è questo: la traiettoria attuale potrebbe portarci in quella direzione prima che istituzioni, imprese e società abbiano strumenti adeguati per governarla.

Il dato più significativo riguarda Claude: a maggio 2026, secondo Anthropic, oltre l'80% del codice integrato nella base di produzione era attribuibile al modello; nello stesso periodo, l'ingegnere tipico integrava circa otto volte più codice al giorno rispetto al 2024. Si leggono questi numeri come il segnale di un passaggio più profondo: l'essere umano non è più sempre il produttore diretto del codice, ma chi definisce l'obiettivo, supervisiona e valuta il risultato. Chi sceglie i problemi davvero importanti, valuta l'affidabilità dei risultati e decide quando abbandonare una linea di ricerca? Per ora, questo giudizio resta il principale vantaggio comparato dell'essere umano.

Se l'IA automatizza soltanto l'esecuzione, siamo di fronte a una grande moltiplicazione della produttività. Se invece comincia ad automatizzare anche il giudizio di ricerca, il rapporto cambia radicalmente: non si tratta più solo di usare uno strumento potente, ma di governare un processo che si ridefinisce dall'interno. La domanda, quindi, non è più solo quanto lavoro possa svolgere l'IA, ma chi capisce e controlla il ciclo che l'IA sta accelerando.

Anche le conseguenze professionali sono ambivalenti. Piccoli gruppi possono ottenere risultati che in passato avrebbero richiesto grandi organizzazioni; allo stesso tempo, possono indebolirsi competenze, apprendistato e senso di padronanza. Lì articolo coglie bene questo aspetto: la produttività può aumentare mentre diminuisce

la percezione di contare davvero. Se il programmatore scrive meno codice e coordina agenti che lo fanno al suo posto, la competenza si sposta verso formulazione dei problemi, verifica e responsabilità.

Sul piano politico l' articolo comunica un dilemma. Rallentare lo sviluppo dell'IA frontier potrebbe essere prudente, perché darebbe più tempo a sicurezza, allineamento e istituzioni. Una pausa unilaterale, però, favorirebbe gli attori meno cauti. Anthropic osserva che un rallentamento credibile richiederebbe coordinamento internazionale, verifiche robuste e fiducia tra laboratori e Stati. Il problema è che i training run sono difficili da osservare e gli incentivi a proseguire in segreto restano molto forti.

La valutazione più equilibrata è duplice. L'articolo non dimostra che il pieno miglioramento ricorsivo sia imminente: molte evidenze provengono da Anthropic e le linee di codice non misurano da sole qualità, sicurezza o comprensione. Allo stesso tempo, sarebbe riduttivo liquidare il tema come fantascienza. Le prove disponibili indicano che l'IA sta già entrando nel ciclo produttivo dell'IA stessa e che il ruolo umano si sta concentrando sempre di più su direzione, verifica e responsabilità.

La questione decisiva, in conclusione, non è se l'IA possa scrivere più codice di un essere umano: in alcuni contesti avanzati questo accade già. La vera domanda è se società democratiche, istituzioni scientifiche e imprese sapranno governare una tecnologia che accelera se stessa. In questo senso, When AI builds itself non è una profezia, ma un avvertimento. Il futuro dell'IA dipenderà meno dalla sola potenza dei modelli e più dalla qualità delle regole, delle verifiche e delle responsabilità che si sapranno costruire.

Luciano Saporito