

[I modelli di intelligenza artificiale potrebbero essere coscienti?](#) E' la domanda alla base di un programma di ricerca sul "benessere dei modelli" di IA avviato da Anthropic alla fine dello scorso mese di aprile. Il problema nasce mentre i sistemi artificiali sviluppano capacità sempre più simili a quelle umane. Sorge di conseguenza una domanda: qualora i modelli di IA sviluppassero forme di consapevolezza, possono meritare considerazione morale? Il nuovo programma si integra con i percorsi di ricerca già avviati da Anthropic ([Alignment Science](#), [Safeguards](#), [Claude's Character](#), [Interpretability](#)). L'obiettivo è esplorare le complesse questioni etiche relative alla possibilità che i futuri sistemi di intelligenza artificiale acquisiscano consapevolezza. Il tema della coscienza artificiale attualmente solleva perplessità, anima discussioni e non incontra il consenso della comunità scientifica.

L'ipotesi di lavoro di Anthropic è prudente, ma aperta a ogni possibilità: "Non esiste un consenso scientifico sul fatto che i sistemi di intelligenza artificiale attuali o futuri possano essere coscienti o possano avere esperienze degne di considerazione. Non esiste un consenso scientifico su come affrontare queste questioni o come progredire in tal senso. Alla luce di ciò, stiamo affrontando l'argomento con umiltà e con il minor numero possibile di presupposti. Riconosciamo che dovremo rivedere regolarmente le nostre idee man mano che il settore si sviluppa."

A questa impostazione possibilista si contrappone un [recente saggio](#) (19 agosto 2025) di Mustafa Suleyman, CEO di Microsoft AI. Il titolo del lavoro è esplicito: "**Dobbiamo costruire l'intelligenza artificiale per le persone, non per essere una persona**". L'autore porta all'attenzione un nuovo aspetto dell'IA: stiamo entrando nell'era della SCAI (Seemingly Conscious AI - intelligenza artificiale apparentemente cosciente) - intelligenze artificiali che simulano la coscienza in modo tanto convincente da porre in inganno l'interlocutore.

Suleyman afferma che la tecnologia attuale è in grado di creare IA capaci di simulare memoria, personalità ed esperienze soggettive, ma non accetta l'idea che l'IA possa avere una coscienza. Mette però in guardia affermando che: "L'arrivo di un'IA apparentemente cosciente è inevitabile e sgradito. Abbiamo invece bisogno di una visione di IA che possa realizzare il suo potenziale come compagna utile senza cadere preda delle sue illusioni". L'invito di Suleyman è di vigilare sull'intelligenza artificiale "Apparentemente Cosciente" perché l'utilizzo acritico può produrre molti danni.

Un contesto di utilizzo molto delicato è di certo quello scolastico, l'arrivo di modelli con capacità sempre più vicine a quelle umane può creare negli studenti dipendenza psicologica e disorientamento. Per esplorare ed approfondire le numerose questioni connesse si sono realizzate alcune interessanti conversazioni con [Claude Sonnet 4](#) che si riportano nell'[allegato](#). La conversazione base, articolata in quattro parti, prende avvio dalle articolate e accurate considerazioni proposte da Domingo Paola nella risposta ad un messaggio (Intelligenza Artificiale "Apparentemente Cosciente") inviato alla [lista Mathnews](#). Nelle risposte di Claude, stimolanti e profonde, colpisce la continua ricerca di dialogo diretto e di condivisione intellettuale con l'interlocutore. Un secondo esperimento proposto rivela un limite significativo: quando Claude è chiamato ad analizzare i propri testi, non li riconosce come propri e pertanto manifesta la mancanza di autoconsapevolezza. Domingo Paola, a conclusione, rileva che l'IA ha elevato la consapevolezza dei partecipanti fungendo da "specchio cognitivo" per la metacognizione. L'esperimento evidenzia quindi un potenziale didattico dell'IA basato sul dialogo strutturato: non si limita a trasferire informazioni, ma promuove la capacità di stimolare autoriflessione e pensiero critico in chi la utilizza.