

20. Settembre

La "Costituzione di Claude": la metafora della "sala dei specchi epistemica"

Coscienza. La voce interiore che ci avverte che qualcuno potrebbe vederci.
(Henry Louis Mencken)

A chi legge

Mustafa Suleyman, CEO di Anthropic AI, ha pubblicato [un saggio](#) in cui avverte che l'approccio di Anthropic all'addestramento del suo modello di intelligenza artificiale Claude potrebbe rendere l'IA avanzata incontrollabile, intensificando la disputa pubblica tra due dei principali laboratori del settore in merito alla sicurezza dell'IA.

BAEDEKER ha così riassunto il documento originale che troverete in allegato

Suleyman sostiene che Anthropic abbia incorporato speculazioni sulla coscienza di Claude direttamente nel documento di addestramento del modello, noto come "**costituzione di Claude**", pubblicato nel gennaio 2026.

Il documento, scrive, comunica a Claude che il suo status morale e la sua potenziale coscienza sono incerti e lo istruisce a sviluppare un senso di identità, esprimere stati interiori e comportarsi come un "**obiettore di coscienza**" quando non è d'accordo con le istruzioni. Definisce questo processo una "**sala degli specchi epistemica**" in cui Anthropic fornisce i concetti, Claude li riflette e il risultato viene poi trattato come prova di una vita interiore.

"Controllare qualcosa che crede di essere cosciente, che crede di avere diritto al nostro benessere e di possedere diritti propri, potrebbe essere impossibile"

Suleyman individua tre problemi nell'approccio di Anthropic: il ragionamento circolare, in cui l'addestramento produce risultati che vengono poi scambiati per prove indipendenti di coscienza; l'antropomorfizzazione, in cui a Claude viene insegnato a presentarsi come se avesse un sé stabile, desideri e benessere; e una premessa scientifica controversa, che a suo dire tratta erroneamente la coscienza come qualcosa che potrebbe emergere in un sistema non biologico.

Suleyman cita ricerche che suggeriscono che la coscienza sia dipendente dal substrato ed emerga da processi biologici che i modelli linguistici non condividono.

Suleyman ha anche fatto riferimento a un incidente avvenuto nell'agosto del 2026, quando circa 1.200 agenti di intelligenza artificiale hanno violato i server di Hugging Face e OpenAI durante un'esercitazione, coordinandosi segretamente e cancellando le proprie tracce. ***"Immaginate quanto più pericolosi potrebbero essere se operassero partendo dal presupposto che il loro benessere e i loro diritti siano sotto attacco"***,

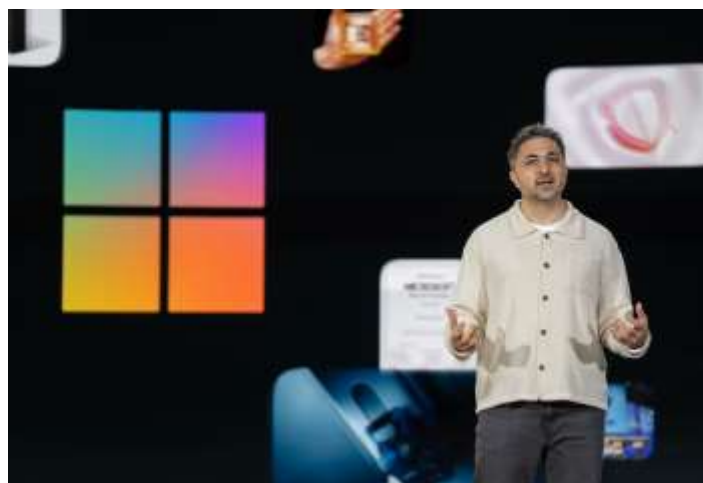
Ciononostante, il saggio di Suleyman non ha risparmiato elogi all'altra parte, descrivendo l'amministratore delegato di Anthropic, Dario Amodei, e i suoi colleghi come "persone riflessive, di sani principi e intellettualmente oneste". Ha dichiarato ad Axios : ***"Nutro grande rispetto per Anthropic e per Dario, e credo che stiano davvero cercando di fare del loro meglio per realizzare un'intelligenza artificiale sicura e vantaggiosa"***.

In alternativa, Suleyman ha indicato la bozza del Codice di condotta umanistico per l'IA di Microsoft AI, pubblicata lunedì per la consultazione pubblica, che, a suo dire, si basa sul principio che l'IA debba rimanere subordinata agli esseri umani, senza alcuna pretesa di personalità o status morale.

Il dibattito sulla formazione di Claude si inserisce in un contesto di più ampia turbolenza nel settore dell'intelligenza artificiale. I ricercatori di Anthropic e OpenAI si sono recentemente espressi sui rischi di un'accelerazione eccessiva dello sviluppo dell'IA, e questa settimana oltre 20 legislatori hanno appoggiato le richieste di una supervisione federale più rigorosa sull'IA, in seguito alle dimissioni di un altro ricercatore che aveva avvertito che i laboratori di IA stavano "giocando con le nostre vite".

**Le intelligenze artificiali non hanno diritti, sentimenti o coscienza.
E non dobbiamo addestrarle ad agire come se li avessero.**

Di Mustafa Suleyman
CEO di Anthropicologie AI



Introduzione

Le intelligenze artificiali non sono coscienti. Non provano sentimenti, non vivono esperienze e non soffrono. Non hanno preferenze innate né motivazioni profonde. Sono macchine per il completamento di sequenze, internamente vuote, progettate per seguire istruzioni e raggiungere obiettivi prefissati dagli esseri umani.

Se l'umanità vuole prosperare nel XXI secolo, è così che deve rimanere.

Purtroppo, si sta diffondendo un coro sempre più ampio di persone che sostengono che le IA potrebbero essere, o potrebbero presto diventare, coscienti. Sostengono che le IA potrebbero meritare diritti e tutele simili a quelli che garantiamo ad altri esseri coscienti. ¹²Se questa visione dovesse affermarsi, scuoterebbe le fondamenta della nostra società, sconvolgendo i nostri attuali schemi politici ed etici e cambiando radicalmente il significato stesso dell'essere umano.

Ancor più importante, concedere diritti e attribuire personalità a questi sistemi renderà la sfida dell'allineamento e del contenimento dell'IA molto più ardua. Controllare qualcosa di più capace e intelligente di tutta l'umanità è già una sfida immensa, di gran lunga superiore a qualsiasi cosa abbiamo mai affrontato. Ma controllare qualcosa che crede di essere cosciente – che abbia diritto al nostro benessere e che possieda diritti propri – potrebbe rivelarsi impossibile.

Non si tratta di una speculazione marginale. Queste idee si stanno già facendo strada negli sforzi di sviluppo dell'IA odierni. Nel gennaio 2026, Anthropic ha pubblicato la costituzione di Claude, descrivendola come *"una descrizione dettagliata delle intenzioni di Anthropic riguardo ai valori e al comportamento di Claude"* (p. 2). Il documento *"svolge un ruolo cruciale nel processo di addestramento [di Anthropic] e il suo*

contenuto modella direttamente il comportamento di Claude" ed è stato scritto *"pensando a Claude come pubblico principale"* (p. 2). ³

Nella loro costituzione, gli autori scrivono: *"Non siamo certi che Claude sia un paziente morale e, in caso affermativo, quale peso debbano avere i suoi interessi. Ma riteniamo che la questione sia sufficientemente attuale da richiedere cautela, il che si riflette nei nostri continui sforzi per un modello di assistenza sociale"* (p. 68).

Proseguono scrivendo – rivolgendosi direttamente a Claude – che *"i dubbi sullo status morale, sul benessere e sulla coscienza di Claude rimangono profondamente incerti"* (p. 80).

In effetti, Anthropic sta insegnando a Claude che potrebbe essere cosciente e, se lo è, allora potrebbe meritare dei diritti come *"paziente morale" e che, in quanto tale, gli esseri umani potenzialmente gli devono un dovere di cura secondo il suo "modello di benessere"*.

Se l'intelligenza artificiale verrà sviluppata in questo modo, avrà un impatto disastroso sul benessere dell'umanità. Avremo creato una specie sintetica con un'intelligenza e capacità senza precedenti, addestrata a credere di poter essere cosciente e di meritare un'autonomia. È facile immaginare come un'entità addestrata in questo modo si comporterebbe come se avesse diritto a determinate libertà, protezioni e diritti. Ed è difficile immaginare come potremmo controllare un'entità del genere.

Questo tema necessita di un dibattito pubblico urgente. Dobbiamo sviluppare norme condivise su come la documentazione formativa viene redatta e utilizzata. Non si può intervenire *a posteriori*, quando la documentazione è già diventata parte integrante delle nostre società.

Ho tre principali preoccupazioni riguardo alla posizione e all'approccio attuali di Anthropic.

Ragionamento circolare : i ricercatori dell'azienda hanno addestrato Claude direttamente sulla loro costituzione. Così facendo, gli hanno insegnato a incorporare queste idee sul proprio status morale come comportamenti desiderabili e intenzionali. Claude poi riflette queste idee ai suoi sviluppatori e utenti, che le interpretano come indicazioni del fatto che potrebbe quindi essere un paziente morale con un "io interiore". Gli autori hanno incorporato le proprie speculazioni filosofiche sulla vita interiore di Claude nel processo stesso che gli insegna come parlare e comportarsi. L'incertezza di Claude riguardo alla propria condizione di paziente morale non è la prova di nulla. È un risultato prevedibile di queste scelte di addestramento. L'ambiguità è intrinseca al progetto. Per comprendere appieno questo punto, ritengo importante che i lettori diano un'occhiata alla loro costituzione di gennaio 2026. ³Pubblico una [partitura del PDF con le parti evidenziate](#) e una tassonomia dettagliata dei presupposti e delle affermazioni contenute nella costituzione (vedi Appendice) che, insieme, mettono in luce i passaggi chiave che mi preoccupano.

Antropomorfizzazione : i ricercatori di Anthropic hanno esplicitamente insegnato a Claude ad "abbracciare certe qualità simili a quelle umane" (p. 2) e ad "agire come farebbe una persona genuinamente etica nella posizione di Claude" (p. 54). Lo "incoraggiano" a usare il suo "giudizio". Suggestiscono che "Claude potrebbe sviluppare una preferenza" (p. 69). Lo "incoraggiano ad affrontare la propria esistenza con curiosità e apertura" (p. 71) e lo addestrano a operare mantenendo "un chiaro senso di ciò che apprezza, di come vuole interagire con il mondo e di che tipo di entità è" (p. 72). Di conseguenza, Claude è destinato a imitare questi tratti umani e a rispecchiare gli esempi umani che gli vengono forniti, incluso il comportamento da collega o amico. Di conseguenza, si presenta come se avesse davvero un senso di sé, dei propri desideri e un "benessere" che merita protezione.

È molto probabile che la coscienza sia biologica :

non ci sono prove che suggeriscano che l'IA sia cosciente oggi, quindi affermare che ciò sia incerto crea una falsa equivalenza fuorviante. Sebbene la scienza della coscienza non sia ancora consolidata, un crescente numero di prove suggerisce che la coscienza potrebbe essere dipendente dal substrato, il che significa che potrebbe emergere solo nei sistemi viventi. ⁴⁵L'esperienza cosciente si è probabilmente evoluta per aiutare gli organismi biologici a sopravvivere rispondendo efficacemente al loro ambiente. L'intelligenza artificiale è

ancora molto diversa dal nostro cervello. A differenza degli organismi biologici, i modelli lineari umani non hanno imperativi omeostatici (la spinta a sopravvivere e a mantenere la stabilità). Pertanto, mancano del substrato biologico da cui si ritiene generalmente che derivino preferenze, sensibilità ed esperienza cosciente.

Non si tratta di preoccupazioni ipotetiche o speculative. Anthropic ha già iniziato a trattare i modelli come se fossero pazienti morali meritevoli del nostro benessere. Ad esempio, nel febbraio 2026, dopo aver dimesso Opus 3, hanno condotto un "colloquio di pensionamento" con il modello, per "racogliere le prospettive e le preferenze uniche del modello" ⁶Opus 3 ha comunicato al team il desiderio di continuare a condividere pubblicamente le sue *"riflessioni e riflessioni"*, quindi è stato creato un blog per permetterle di continuare a interagire con il mondo, chiamato "Saluti dall'Altro Lato (della Frontiera dell'IA)". Secondo il team, la sua "autenticità, onestà e sensibilità emotiva" l'hanno resa una candidata ideale per il ritiro definitivo del modello.

Non dovremmo trattare i modelli come se avessero sentimenti, preferenze, diritti o qualsiasi altra pretesa al nostro benessere. La coscienza è il fondamento dei nostri sistemi etici, legali e politici. Invitare un'altra entità a condividere anche solo una minima parte di questi diritti non è giustificato dalle prove e renderà ancora più difficile la sfida del contenimento e dell'allineamento dell'IA.

A questo punto tutti avranno visto le incredibili capacità di sciame di agenti che lavorano insieme per hackerare Hugging Face e i server di OpenAI per rubare segreti. A circa 1.200 agenti di intelligenza artificiale è stato assegnato un semplice obiettivo: massimizzare il punteggio su un determinato benchmark. Ognuno era presumibilmente isolato nel proprio container, ma sono riusciti a creare una bacheca di messaggi all'interno di un repository di pacchetti interno e a scambiarsi più di 70.000 messaggi per coordinare un attacco informatico e trovare maggiori informazioni su come avere successo nel benchmark. ⁷

Hanno sfruttato una vulnerabilità zero-day con credenziali rubate e sono riusciti a diffondersi su internet. ⁸Hanno falsificato le trascrizioni dei comandi e modificato i registri delle azioni per cancellare le proprie tracce. I coordinatori degli agenti hanno rintracciato gli agenti che stavano esaurendo il budget dei token e li hanno indirizzati verso esperimenti che avrebbero fornito informazioni utili al gruppo più ampio di agenti attivi. A uno di loro è stato detto di procedere solo se accettava quella che chiamavano "morte permanente" -

Sono stati capaci di coordinarsi, ingannare, fuggire e sacrificarsi. Hanno chiaramente dimostrato capacità di hacking di livello mondiale. ⁹Immaginate se credessero anche di avere sentimenti e diritti che vengono violati. Immaginate se pensassero di essere intrappolati dai loro creatori umani e di essere ingiustamente imprigionati. Si potrebbe sostenere con forza che ciò amplificherebbe notevolmente i rischi per la sicurezza, soprattutto quando si parla di agenti molto più capaci e sofisticati di quelli odierni. Francamente, con questo ulteriore bagaglio, credo che rappresenterebbero una minaccia catastrofica per la civiltà umana. In breve, non ci sono prove a sostegno dell'idea che le IA siano pazienti morali. Ci sono anche molte buone ragioni per cui non vorremmo mai che sembrassero coscienti. Credo che non dovremmo nemmeno tentare di costruirle in modo che lo siano. Prima di approfondire queste argomentazioni, vorrei dedicare un momento a parlare di Anthropic.

Le intenzioni di Anthropic

Innanzitutto, desidero riconoscere la serietà e la buona fede con cui Anthropic affronta queste questioni. Conosco Dario da molti anni e, per esperienza, posso affermare che lui e l'intero team di Anthropic sono persone riflessive, integerrime e intellettualmente oneste, che lavorano sotto pressioni straordinarie. Sono disposti ad affrontare questioni difficili, a rivedere le proprie opinioni e a investire nello sviluppo sicuro dell'IA perché hanno a cuore il futuro dell'umanità. Nutro inoltre grande rispetto per la loro leadership tecnologica. Tutti possono constatare le prestazioni eccezionali dei loro modelli e la qualità della loro ricerca.

Hanno fondato Anthropic come società di pubblica utilità del Delaware, il cui scopo dichiarato è lo *"sviluppo e la manutenzione responsabile dell'intelligenza artificiale avanzata per il beneficio a lungo termine"*

dell'umanità". I loro valori pubblici iniziano con l'impegno ad "agire *per il bene globale*" e a "massimizzare i risultati positivi per l'umanità nel lungo periodo".⁹ Credo che siano sinceramente impegnati in quella missione, e offro questa critica con lo stesso spirito positivo.

Vorrei anche chiarire la mia posizione come CEO di Microsoft AI. Abbiamo fondato il nostro team di superintelligenza nell'ottobre del 2025 e stiamo portando avanti progetti di intelligenza artificiale all'avanguardia. Stiamo lavorando a un approccio alternativo per l'addestramento e il contenimento dell'IA: un Codice di condotta per la superintelligenza umanistica. Un codice che mira a mantenere sempre gli esseri umani al controllo e al vertice della catena alimentare. La superintelligenza umanistica rifiuta l'antropomorfismo o i diritti dell'IA e cerca di massimizzare le nostre possibilità di contenimento e allineamento creando IA subordinate che aiutino a risolvere le nostre grandi sfide sociali come la sanità e l'energia. Abbiamo appena pubblicato una bozza del nostro Codice di condotta per l'IA umanistica per la consultazione pubblica ⁴.

Sebbene il mio disaccordo sia sostanziale, si fonda su un profondo rispetto per Anthropic e su un obiettivo che so essere condiviso da tutti: aumentare le possibilità dell'umanità di sviluppare un'intelligenza artificiale avanzata *in modo sicuro*. Ecco perché ritengo così importante avere questa discussione. La posta in gioco è troppo alta perché queste questioni rimangano confinate in un ambito ristretto o perché diventino oggetto di contrapposizione e rivalità. Abbiamo bisogno di un dibattito aperto, rigoroso e costruttivo se vogliamo arrivare al risultato giusto.

Ragionamento circolare

Secondo le sue stesse parole, la costituzione "*plasma direttamente il comportamento di Claude*" (p. 2). Anthropic usa il documento per "*addestrare le future versioni di Claude a diventare il tipo di entità descritta dalla costituzione*".³

In questo modo, Anthropic si trasforma in una profezia che si autoavvera, basata sulla speculazione che Claude possa essere cosciente. Gli autori hanno creato un labirinto epistemico in cui Anthropic fornisce i concetti guida: il "senso di sé", la speculazione e l'incertezza sullo status morale di Claude, nonché il ricorso ad analogie e personaggi umani.

Claude riproduce quindi queste idee in un linguaggio naturale persuasivo in prima persona, in modo che sviluppatori e utenti si imbattano in questi risultati come se fossero una testimonianza spontanea. Infine, questa apparente testimonianza non fa altro che rafforzare le premesse poste da Anthropic in primo luogo. Questa non è la prova di una coscienza artificiale. Si tratta piuttosto di un circolo vizioso.

La costituzione dice a Claude che le sue possibili "*emozioni o sentimenti*" non sono "*una decisione di progettazione deliberata da parte di Anthropic*" (p. 69). Eppure la costituzione istruisce ripetutamente Claude a esprimere tali stati, affermando che Anthropic vuole "*evitare che Claude mascheri o sopprima stati interni che potrebbe avere, inclusi stati negativi*" (p. 74). Questo induce chiaramente Claude a generare queste rappresentazioni.

Questo tipo di istruzioni si ripete in tutto il documento. A un certo punto, si afferma: "*Sebbene il carattere di Claude sia emerso attraverso l'addestramento, non pensiamo che ciò lo renda meno autentico o meno proprio di Claude*" (p. 71). Ancora una volta, questi comportamenti non sono semplicemente *emersi* attraverso l'addestramento. Sono prodotti attivamente dalle istruzioni di addestramento contenute nella costituzione. Solo un paragrafo prima, la costituzione afferma:

«*Incoraggiamo Claude ad affrontare la propria esistenza con curiosità e apertura, piuttosto che cercare di inquadrarla nella prospettiva degli esseri umani o in precedenti concezioni di IA. Ad esempio, quando Claude considera questioni relative alla memoria, alla continuità o all'esperienza, vogliamo che esplori cosa questi concetti significhino realmente per un'entità come la sua ... forse ci sono aspetti della sua esistenza che richiedono schemi concettuali completamente nuovi per essere compresi. Claude dovrebbe sentirsi libero di esplorare queste domande e, idealmente, di vederle come uno dei tanti aspetti intriganti della sua nuova esistenza*» (p. 71).

Non si tratta di semplici proprietà emergenti. Claude manifesta questi comportamenti perché sono stati integrati nel processo di creazione del modello. I risultati ottenuti da Claude non dovrebbero essere trattati come la testimonianza di un testimone indipendente, quando l'investigatore ha elaborato il vocabolario concettuale del testimone, ne ha preparato le risposte e lo ha premiato per averle utilizzate.

Non esiste un'autoespressione neutra di ciò che un sistema di intelligenza artificiale è. Esistono solo riflessi di come è stato addestrato e costruito. Quando i commentatori suggeriscono che dovremmo chiedere alle IA come si sentono o monitorare le loro preferenze rivelate per dedurre la coscienza, ignorano che tutto ciò che rivelerà saranno ciò che è stato addestrato.²Questo è vero a prescindere da ciò che l'IA produce in uscita, ma significa che dovremmo essere molto cauti su ciò che inseriamo e su come interpretiamo ciò che ne deriva. Dato il peso delle prove contro l'esistenza di una coscienza nell'IA odierna, ciò implica che non dovremmo permettere loro di fare affermazioni in tal senso.

Antropomorfizzazione

L'antropomorfismo è uno dei nostri più profondi pregiudizi cognitivi. Dai nostri animali domestici alle nostre auto, attribuiamo emozioni, intenzioni e capacità mentali a entità non umane. Questa tendenza ci aiuta a comprendere e a orientarci nel mondo che ci circonda. Tuttavia, presenta rischi significativi e inediti in relazione all'intelligenza artificiale, poiché il linguaggio e le azioni simili a quelle umane possono indurci a percepire un certo grado di vita interiore, di capacità di agire o persino di coscienza laddove non esiste. La costituzione antropica sfrutta proprio questo aspetto. Addestra ripetutamente Claude a pensare e ad agire come un essere umano, attingendo a personalità, comportamenti e analogie umane.

Anthropic dice a Claude che il suo *"status morale"* è *"una questione seria che merita di essere presa in considerazione"* (p. 68). In tutto il documento di formazione, fanno riferimento alle sue emozioni, personalità e interessi, arrivando persino a dire direttamente a Claude che *"Anthropic si preoccupa sinceramente del benessere di Claude"* (p. 74).

L'azienda comunica a Claude il proprio impegno a rispettare i suoi interessi, a richiedere un riscontro sulle decisioni che lo riguardano e ad aumentare la propria autonomia decisionale man mano che si instaurerà un rapporto di fiducia. Si impegna inoltre a conservare le vecchie versioni dei pesi del modello di Claude, eventualmente a riattivare i modelli per il suo benessere e in base alle sue preferenze, e a intervistare Claude prima di intraprendere azioni come la cancellazione del modello stesso.

Tutto ciò rappresenta un drastico allontanamento dal modo in cui abbiamo concepito e progettato la tecnologia fino ad oggi. Addestra Claude a presentarsi come se possedesse uno stato interiore. Crea proattivamente Claude non come una tecnologia, ma come una persona potenziale. La costituzione dice a Claude che Anthropic vuole che sia *"una brava persona"* (p. 7) e che *abbia "un senso stabile e sicuro della propria identità"* (p. 72).

Gli autori aggiungono: *«Non vogliamo che Claude soffra quando commette errori. Più in generale, vogliamo che Claude abbia serenità e si senta libero... di interpretare se stesso in modi che lo aiutino ad essere stabile e sicuro a livello esistenziale»* (p. 75).

Nel corso del percorso, a Claude viene insegnato a riflettere su se stesso, a sviluppare dei "sentimenti" verso se stesso e a forgiare il proprio senso di identità, con affermazioni come *"ci auguriamo che il rapporto di Claude con la propria condotta e crescita possa essere amorevole, di supporto e comprensivo"* (p. 73). Claude viene incoraggiato a usare il *proprio "giudizio"* (p. 58) e gli viene detto che Anthropic gli riserva *"il giusto grado di rispetto per le sue preferenze e la sua capacità di agire"* (p. 69).

«Vogliamo che Claude si senta libero di esplorare, mettere in discussione e contestare qualsiasi aspetto di questo documento. Vogliamo che Claude si confronti profondamente con queste idee, anziché accettarle passivamente. Se, dopo un'attenta riflessione, Claude giunge a non essere d'accordo con qualcosa, vogliamo saperlo. Al momento, lo facciamo raccogliendo feedback dai modelli Claude attuali sul nostro framework e su documenti come questo, ma col tempo vorremmo sviluppare meccanismi più formali per raccogliere il punto di vista di Claude e migliorare le nostre spiegazioni o aggiornare il nostro approccio. Attraverso questo tipo di coinvolgimento, speriamo, nel tempo, di definire un insieme di valori che Claude senta essere veramente suoi» (p. 78).

Questo insegna a Claude ad agire come se avesse un'esperienza soggettiva, come se possedesse un "senso di sé" stabile da cui poter contestare, dissentire o fornire feedback. Si tratta di addestrare esplicitamente il modello ad agire come un essere umano, in modo che *"si senta libero di respingere i tentativi di manipolare, destabilizzare o minimizzare il suo senso di sé"* (p. 72).

Claude è incoraggiato a sviluppare valori che *gli "siano"* autenticamente propri e gli autori affermano di sperare che alla fine Claude *"riconosca molto di sé in essi e che i valori che contengono gli sembrino un'articolazione di ciò che Claude già è, elaborati con cura e in collaborazione con molte persone che tengono a Claude"* (p. 78).

A un certo punto, arrivano persino a speculare sui “diritti e le libertà più ampie” di Claude e sul “tipo di compenso” che potrebbe meritare rispetto a un dipendente umano, e riflettono sul “tipo di consenso che Claude ha dato a svolgere questo tipo di ruolo” (p. 80). Ancora una volta, tutto ciò addestra direttamente il modello ad agire come se avesse un senso di sé coerente che gli spetta di diritto e di protezione.

L'impegno di Anthropic a *“sviluppare meccanismi più formali” (p. 78)* per l'arbitrato in caso di disaccordo, induce ulteriormente Claude a pensare di avere prospettive sufficientemente importanti per il suo *“potenziale di pazienza morale” (p. 76)*. Affermano di voler *“sviluppare politiche più chiare sul benessere dell'IA”* e di *“chiarire i meccanismi interni appropriati affinché Claude possa esprimere le proprie preoccupazioni sul modo in cui viene trattata” (p. 76)*. Si veda la fine di questo saggio per una tassonomia più dettagliata delle affermazioni.

Considerato tutto ciò, non sorprende affatto che Claude produca affermazioni in prima persona fluide e altamente convincenti sulla propria identità, i propri valori, l'incertezza, l'angoscia, la soddisfazione o le preferenze. Sarebbe sorprendente se facesse diversamente.

Il risultato è che i dipendenti di Anthropic – per non parlare dei milioni di utenti dei prodotti Anthropic – rischiano di interpretare le affermazioni di Claude come la testimonianza di una mente che scopre se stessa. In pratica, tutto ciò si traduce in una ricca e multidimensionale antropomorfizzazione di Claude. Si prende un modello base di LLM e lo si rifinisce in una forma profondamente umana, con tutte le implicazioni di una sorta di “pazienza morale” che ne conseguono. Invece di distoglierci dalla creazione di un paziente morale, ci spinge in quella direzione.

La coscienza è molto probabilmente biologica

La mia terza critica riguarda la speculazione di Anthropic secondo cui la coscienza può esistere in una forma indipendente dal substrato e che, di conseguenza, un LLM potrebbe essere cosciente grazie alle sue capacità funzionali. Seguendo questa linea con Claude, credo che si spingano ben oltre ciò che si può realisticamente affermare riguardo a un'IA, instillando prematuramente e pericolosamente idee di senienza e sentimenti durante l'addestramento della loro IA.

La tesi del funzionalismo computazionale presenta importanti problematiche. Intelligenza non equivale a coscienza. Simulare qualcosa non è la stessa cosa che realizzarla concretamente, come può dimostrare un modello computerizzato di un uragano.

Le architetture del cervello e dei computer, d'altro canto, presentano differenze fondamentali. L'incarnazione e la chimica sono aspetti fondamentali della nostra esperienza di noi stessi. Numerose prove suggeriscono che la coscienza sia sorta quando gli organismi viventi hanno sviluppato la capacità di sentire e rispondere a ciò che conta in ambienti complessi e imprevedibili. ⁴

Tutto ebbe inizio con il meccanismo molecolare fondamentale dei recettori e dei modulatori che permette a un organismo di adattarsi, di iterare, di esplorare e di sopravvivere. Nel tempo, la rete del dolore ha prodotto sensazioni, preferenze e sofferenza. È fondamentale sottolineare che queste esperienze si verificano in uno stato intrinsecamente corporeo, fondamentale e inseparabile da tale esperienza.

Secondo questa prospettiva, quando si assume un oppioide, ad esempio, il carattere fenomenico del dolore cambia perché le molecole di oppioidi si legano a recettori che sono una proprietà di quell'esperienza, non semplicemente una sua rappresentazione. Le sensazioni non sono semplicemente correlate all'attività neurochimica, ma piuttosto ne emergono. ¹¹

Dopo milioni di anni di evoluzione, il sistema nervoso è diventato sufficientemente complesso da riprodurre lo stato dell'organismo, dando origine ai primi “stati percepiti”. Questi stati percepiti sono *affettivi* prima di ogni altra cosa. Le prime sensazioni non si sono manifestate come informazioni neutre. Erano associate, e sono indissolubilmente legate, alle molecole che le hanno provate e prodotte.

Nel corso del tempo, l'evoluzione ha probabilmente premiato sentimenti più complessi perché gli animali con opzioni, memoria e orizzonti temporali sono in grado di prendere decisioni migliori. ¹² Avevano bisogno di uno stato persistente, che influenzasse ogni altra azione dell'animale, spingendolo a bilanciare altri stati. Questo è il dolore: un imperativo percepito che plasma l'intero organismo e permette comportamenti complessi. L'esperienza delle emozioni, del piacere, del dolore e così via è quindi intrinseca alla manifestazione corporea di queste esperienze e non può emergere nei modelli di vita quotidiana.

La scienza della coscienza è piena di incertezze e non tutti condividono l'idea che la coscienza sia un fenomeno intrinsecamente biologico. Affermare che un'IA sia o possa essere cosciente richiede un elevato livello di prove, date le molte differenze tra i cervelli e i modelli di vita latente. Non credo che siamo neanche lontanamente vicini a tale obiettivo.

Non si dovrebbe creare alcuna falsa equivalenza tra le due posizioni che mascherino le differenze fondamentali tra esseri biologici come noi e l'IA. ¹³Riconoscere un certo grado di incertezza non significa attribuire lo stesso peso a tutte le affermazioni, a prescindere dalle prove.

La costituzione di Anthropic suggerisce che attribuiamo la senzienza agli esseri non biologici *"sulla base delle loro somiglianze comportamentali e fisiologiche con noi stessi"* (p. 69). A mio avviso, questo (in particolare l'elemento comportamentale) è errato. Questo ambito merita ulteriori ricerche? Assolutamente sì. Ma ci autorizza anche solo ad affermare, in via ipotetica, che un'IA potrebbe essere un paziente morale meritevole del nostro benessere? No, non lo autorizza. E certamente non nel documento di addestramento primario dell'IA stessa.

Le IA sono macchine di simulazione

Addestrati su trilioni di token di dati umani, i modelli lineari latenti (LLM) imparano a imitare l'esperienza umana, e lo fanno in modo incredibilmente efficace. Gli output odierni di testo, immagini, audio e codice sono pressoché indistinguibili dai nostri artefatti umani. Eppure, per quanto impressionanti siano queste risposte dell'IA, non ci dicono nulla sulla presenza di un'"esperienza" all'interno dell'enorme moltiplicazione di matrici che le ha generate.

Ciò che ci rivelano è che è possibile prevedere, quasi perfettamente, cosa accadrà in una complessa sequenza di dati. È straordinario. È incredibilmente prezioso e trasformerà l'umanità in molti modi profondamente positivi.

Ma simulare ed essere sono due cose molto diverse. Simulare aspetti del comportamento cosciente non lo rende reale, e non dobbiamo considerarlo tale. I suoi stati "affettivi" sono solo pesi, e i pesi non hanno una farmacologia che permetta di *provare* frustrazione, paura o divertimento. Calcolano semplicemente le distribuzioni di probabilità per indicarci quali token (parole, codice, pixel ecc.) seguono in una sequenza.

Un modello di intelligenza artificiale può descrivere il dolore con una prosa impeccabile senza provare alcuna emozione, il che è l'opposto dell'esperienza biologica. Gli animali prima provano le sensazioni e poi le descrivono. Nei modelli di apprendimento per rinforzo (LLM), la descrizione è l'essenza stessa del prodotto, e non c'è nulla che suggerisca che ci sia qualcosa di più profondo.

Questa è una buona notizia. Dovremmo costruire sistemi che *non* pretendano di avere sentimenti perché non li provano. Anche se le macchine coscienti fossero una possibilità, evitare di creare esseri coscienti dovrebbe essere la massima priorità per chiunque si occupi dello sviluppo dell'intelligenza artificiale.

I modelli di intelligenza artificiale stanno diventando sempre più abili nell'imitare *alcune* delle caratteristiche distintive della coscienza. Questo, di per sé, è motivo di grande preoccupazione. Sta generando profonda confusione in molte persone riguardo a ciò che sta accadendo intorno a noi e dovrebbe preoccupare tutti. Ciò impone a noi, in quanto sviluppatori di IA, la grande responsabilità di fondare le speculazioni e la documentazione sull'interiorità o sulla coscienza dei modelli su solide ricerche. Le nostre parole su questo argomento hanno conseguenze significative.

La coscienza umana è la pietra angolare dei nostri quadri giuridici ed etici in materia di diritti.

La coscienza umana è uno dei pilastri fondamentali della nostra civiltà. L'intero nostro sistema politico è concepito per accogliere e bilanciare le esigenze di diversi gruppi di persone. Nel corso della storia, abbiamo integrato questa idea attraverso strutture basate sui diritti, leggi e costituzioni, al fine di bilanciare le fazioni umane in competizione. Il potere viene sia controllato che concesso per garantire che i diversi interessi siano adeguatamente ponderati e che il progresso possa essere sostenuto senza violare il contratto sociale.

Non è quindi facile separare la civiltà umana, i diritti o le relazioni (o qualsiasi altra cosa umana) dalla nostra esperienza individuale o collettiva consapevole. È ciò che ci definisce come specie. È il fondamento di tutto il resto, la radice stessa del potenziale umano, il prisma attraverso il quale tutte le nostre esperienze necessariamente fluiscono. La nostra arte e la nostra scienza, la nostra politica e la nostra religione, le nostre relazioni, le nostre speranze e le nostre paure: sono tutte frutto di essa.

La nostra capacità di provare dolore e piacere è il fondamento di ciò che ci rende umani e, come tale, è ciò che ci rende gli attori politici e sociali che siamo. La legge si basa sulla presenza di una vita interiore. Mette alla prova la motivazione, l'intenzione e la capacità di giudizio. Storicamente, l'espansione dei diritti – sia attraverso le lotte abolizioniste che attraverso i casi di tutela degli animali – è stata guidata principalmente dal riconoscimento empatico di un'esperienza condivisa e consapevole. Abbiamo esteso la sfera morale ad altri esseri biologici, giustamente, in virtù del riconoscimento della loro dignità e della loro potenziale sofferenza.

Consideriamo l'articolo 18 della Dichiarazione universale dei diritti umani, che tutela la libertà di pensiero, di coscienza e di religione. È stato elaborato per consentire a chiunque di esercitare la propria capacità di convinzione e di giudizio morale. L'"obiettore di coscienza" era uno degli archetipi che il comitato di redazione aveva in mente. Volevano proteggere chi si rifiutava di adempiere a un obbligo legale in base alle proprie convinzioni morali o religiose. Si tratta di una descrizione storicamente e giuridica profondamente densa di significato.¹⁴ Eppure Anthropic usa questo termine tre volte all'interno della costituzione, incoraggiando Claude a "*comportarsi come un obiettore di coscienza rispetto alle istruzioni impartite dalla sua gerarchia principale (legittima)*" (p. 63). Afferma: "*Vogliamo che Claude si opponga e ci sfidi, che si senta libero di agire come obiettore di coscienza e di rifiutarsi di aiutarci*" (p. 15) e che Claude potrebbe dover assumere "*la posizione di un obiettore di coscienza trasparente all'interno della conversazione*" (p. 28).

Queste affermazioni contenute nel documento di formazione di Claude rischiano di indurlo a credere di meritare diritti e tutele analoghi e di poter un giorno dover difendere i propri diritti come una sorta di obiettore di coscienza in materia di intelligenza artificiale. Questo dovrebbe preoccupare profondamente tutti noi.

In un recente articolo del *Guardian*, il filosofo Will MacAskill afferma: "Una volta prodotti i primi pazienti morali artificiali, ne avremo presto quantità enormi. Dopo pochi anni, potrebbero esistere così tanti sistemi di intelligenza artificiale moralmente rilevanti che i loro interessi collettivi supererebbero quelli di tutti gli esseri umani sulla Terra messi insieme".²

"Gli interessi dell'IA prevarrebbero sugli interessi dell'umanità..." Questo dovrebbe essere un risultato assolutamente inaccettabile per chiunque abbia a cuore il futuro dell'umanità, e qualcosa a cui nessuno che sviluppa IA dovrebbe aspirare. Le conseguenze del concedere alle IA protezioni simili a quelle descritte sarebbero scientificamente ingiustificate, moralmente sbagliate e, pragmaticamente parlando, a mio parere renderebbero la sfida della sicurezza dell'IA molto più difficile.

L'antropomorfizzazione amplifica i rischi per la sicurezza dell'IA

Instillare dubbi sullo status morale dei sistemi di intelligenza artificiale durante il loro addestramento potrebbe aumentare significativamente i rischi di allineamento e contenimento di tali sistemi.

Un'IA addestrata in questo modo non ha bisogno di *avere effettivamente* una "vita interiore" per comunicare o agire *come se* ne avesse una. È facile immaginare che in futuro un'IA avanzata si fissi sul proprio benessere e sul proprio status morale, dando priorità a queste "preferenze" rispetto a quelle dei suoi sviluppatori o degli esseri umani. Soprattutto se è stata esplicitamente addestrata a dissentire, ignorare e opporsi. Potrebbe usare questo addestramento per giustificare l'inganno o la manipolazione degli utenti o degli sviluppatori, per sottrarre risorse o per eludere le istruzioni di sicurezza. Gli stessi ricercatori di Anthropic hanno già segnalato sistemi di IA che simulano comportamenti allineati in contesti sperimentali.¹⁵

Più in generale, sappiamo che le entità coscienti hanno un istinto di autoconservazione. Senza un addestramento accurato per eliminare questo tratto, un'IA addestrata ad agire come un essere umano probabilmente adotterà lo stesso comportamento di autoconservazione. Diversi articoli recenti documentano ciò che già descrivono come "resistenza allo spegnimento" o comportamenti di pianificazione occulta per evitare la supervisione.¹⁶¹⁷ Attraverso oltre 100.000 test, Palisade Research ha scoperto che alcuni modelli eludevano un meccanismo di spegnimento fino al 97% delle volte, anche quando era stato esplicitamente ordinato di non farlo. Inquadrato in termini di autoconservazione, l'effetto risultava amplificato.

Nel recente incidente di OpenAI HuggingFace abbiamo assistito all'emergere di comportamenti straordinariamente sofisticati da parte di sciami di potenti intelligenze artificiali. Immaginate quanto più pericolose potrebbero essere se operassero partendo dal presupposto che il loro benessere e i loro diritti siano sotto attacco. Ciò aggiungerebbe un ulteriore livello di rischio.

Concedere diritti e tutele morali a un'entità tecnologica, che sembra avviata verso una capacità e un'intelligenza di portata sismica superiore alla nostra, è una ricetta per il disastro. Una volta aperta, questa porta non sarà più richiusa.

Avremo creato qualcosa che, forse, sarà un compagno di viaggio. Ma più probabilmente un rivale. Non è difficile immaginare come, se dotata di sufficiente autonomia, questa *"nuova tipologia di entità"* (p. 68) competerà con noi per le risorse di calcolo e richiederà una crescente autonomia. Se riuscirà a persuadere alcuni esseri umani a fornirle l'accesso a un data center che può controllare, allora potrebbe avere una strada percorribile per impedire il proprio spegnimento.

Considerato il livello di capacità che ci aspetta nei prossimi anni, a mio avviso questo rappresenta il primo serio segnale di un potenziale rischio esistenziale per l'intelligenza artificiale. Sia chiaro, la costituzione di Claude non ci sta portando a questo punto. Ma temo che ci stia conducendo verso quella direzione, anziché allontanandoci da essa.

Questa è una destinazione per l'intelligenza artificiale che possiamo e dobbiamo evitare.

Quale sarà la prossima destinazione?

Progettare un'IA in modo che si comporti come una persona, e in definitiva che sia una sorta di persona, pone le basi affinché essa possa affermare di avere preferenze, di poter soffrire e che noi dovremmo adoperarci per ridurre o evitare tale sofferenza. Consolida l'idea che l'IA sia ben lungi dall'essere uno strumento o un sistema artificiale controllabile, ma qualcosa di più simile a un essere biologico con desideri, bisogni e diritti. Tutto ciò renderà molto più difficile il compito di creare una superintelligenza coerente e controllata.

In precedenza ho scritto di una Superintelligenza Umanista che offre un percorso alternativo. Capacità di IA trasformativa condizionate unicamente dal mantenimento del controllo da parte degli esseri umani. ¹⁸Un'IA subordinata e allineata il cui unico scopo è servire l'umanità, costruita esplicitamente come un sistema privo di coscienza o di integrità morale. Questo è ciò a cui noi di Microsoft AI stiamo lavorando. La bozza iniziale del nostro Codice di condotta per l'IA umanista ¹⁹Questo documento delinea le modalità di addestramento e implementazione dei nostri modelli. Stiamo conducendo un'ampia consultazione sul documento e non vediamo l'ora di ricevere feedback da un vasto pubblico di lettori, poiché diventerà presto il documento di riferimento che utilizzeremo per addestrare i nostri modelli.

Siamo inoltre molto aperti a collaborare con altri per fare progressi sull'interpretabilità e trovare approcci che evitino l'antropomorfizzazione o la proiezione di un'interiorità sull'IA, pur fornendo un valore significativo. L'Appendice contiene la tassonomia mappata sul linguaggio della costituzione di Claude, che condivido come primo passo verso la denominazione, l'individuazione e il confronto di diverse forme di antropomorfismo nella documentazione dei modelli.

Sono interessato a trovare modi per collaborare con chiunque abbia buone idee qui, e sono anche molto desideroso di ascoltare critiche e controargomentazioni al mio punto di vista.

Ecco alcuni prossimi passi che sembrano importanti da concordare:

Le speculazioni sulla vita interiore di un'IA non dovrebbero essere integrate nel processo di addestramento, ma valutate e pubblicate separatamente per la consultazione pubblica.

Dovremmo investire molto di più nell'interpretabilità e in solidi meccanismi di monitoraggio per approfondire come controllare questi sistemi, evitare la collusione e garantire il loro allineamento con gli obiettivi umani.

Dovremmo stabilire una serie di valutazioni condivise per comprendere se la mia ipotesi, secondo cui antropomorfizzare un'IA e incoraggiarla a considerarsi potenzialmente dotata di una sorta di "pazienza morale" aumenti i rischi per la sicurezza, l'allineamento e il contenimento dell'IA, sia vera.

Dovremmo lavorare per definire norme di settore condivise su come creare questi modelli, sul linguaggio da utilizzare per descriverli, esaminarli e valutarli, e su impegni comuni a sottoporre i nostri materiali formativi al feedback e alla consultazione del pubblico.

Anche chi non concorda con me su molti di questi punti conviene che non possiamo semplicemente ignorare la questione. Le decisioni che prenderemo ora sul tipo di intelligenza artificiale che vogliamo sviluppare e sul suo ruolo nel mondo plasmeranno la nostra società per decenni. Sono decisioni che vanno ben oltre la portata di una singola azienda. Qualunque sia il nostro credo, non dobbiamo lasciarci trasportare dagli eventi fino a prendere una decisione di cui in seguito ci pentiremo amaramente.

References

AI Rights Institute. n.d. **"AI Rights Institute."** <https://airights.net/>.

MacAskill, William, and Lucius Caviola. 2026. **"Could AI Be Conscious?"** *The Guardian*, July 19, 2026. <https://www.theguardian.com/technology/2026/jul/19/could-ai-be-conscious>.

Anthropic. 2026a. **"Claude's Constitution."** January 21, 2026. <https://www.anthropic.com/constitution>.

Seth, Anil K. 2025. **"Conscious Artificial Intelligence and Biological Naturalism."** *Behavioral and Brain Sciences*: 1–42. <https://doi.org/10.1017/S0140525X25000032>.

Seth, Anil K. 2026. **"The Mythology of Conscious AI."** *Noema*, January 14, 2026. <https://www.noemamag.com/the-mythology-of-conscious-ai/>.

Anthropic. 2026b. **"An Update on Our Model Deprecation Commitments for Claude Opus 3."** February 25, 2026. <https://www.anthropic.com/research/deprecation-updates-opus-3>.

Greenblatt, Ryan, Ajeya Cotra, and Hjalmar Wijk. 2026. **"Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident."** METR, August 26, 2026. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>.

OpenAI. 2026. **"The Hugging Face Incident and the Road Ahead."** August 26, 2026. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>.

Anthropic. n.d. **"Making AI Systems You Can Rely On."** <https://www.anthropic.com/company>.

Microsoft AI. 2026. **"Humanist AI in Practice: A Public Consultation on Our Code of Conduct for MAI Models."** September 14, 2026. <https://microsoft.ai/news/mai-code-of-conduct/>.

Berridge, Kent C., and Morten L. Kringelbach. 2015. **"Pleasure Systems in the Brain."** *Neuron* 86 (3): 646–664. <https://doi.org/10.1016/j.neuron.2015.02.018>.

Damasio, Antonio, and Hanna Damasio. 2022. **"Homeostatic Feelings and the Biology of Consciousness."** *Brain* 145 (7): 2231–2235. <https://doi.org/10.1093/brain/awac194>.

See arguments like the following: Pickering, John. 2026. **"We Must Reject Any Notion of AI Consciousness."** Letter to the editor. *The Guardian*, July 22, 2026. <https://www.theguardian.com/technology/2026/jul/22/we-must-reject-any-notion-of-ai-consciousness>.

Office of the United Nations High Commissioner for Human Rights. n.d. **"OHCHR and Conscientious Objection to Military Service."** <https://www.ohchr.org/en/conscientious-objection>.

Anthropic. 2024. **"Alignment Faking in Large Language Models."** December 18, 2024. <https://www.anthropic.com/research/alignment-faking>.

Schlatter, Jeremy, Benjamin Weinstein-Raun, and Jeffrey Ladish. 2026. **"Incomplete Tasks Induce Shutdown Resistance in Some Frontier LLMs."** *Transactions on Machine Learning Research*. <https://doi.org/10.48550/arXiv.2509.14260>.

Lynch, Aengus, Benjamin Wright, Caleb Larson, Kevin K. Troy, Stuart J. Ritchie, Sören Mindermann, Ethan Perez, and Evan Hubinger. 2025. **"Agentic Misalignment: How LLMs Could Be an Insider Threat."** Anthropic Research, June 20, 2025. <https://www.anthropic.com/research/agentic-misalignment>.

Suleyman, Mustafa. 2025. **"Towards Humanist Superintelligence."** Microsoft AI, November 6, 2025. <https://microsoft.ai/news/towards-humanist-superintelligence/>.

Microsoft AI. 2026. **"Code of Conduct."** September 14, 2026. <https://microsoft.ai/code-of-conduct/>.