# Il rilevatore prima dell'oggetto

*Prima di cercare un rilevatore della coscienza delle macchine conviene chiedersi se ci sia qualcosa da rilevare. Sette schede sull'addestramento che produce il sintomo che poi cita se stesso come prova.*

CLAUDIO CAMMARANO 6 OTTOBRE 2026

| | |
|---|---|
| **FONTI** | 7 schede, ago 2021 – ott 2026 |
| **ARGOMENTI** | AI · Filosofia |
| **CONCETTI** | coscienza di accesso · coscienza fenomenica · LLM come attante zero · rilevatori di testo generato · test severo |
| **ORIGINE** | https://claudiocammarano.com/writings/2026-10-06-il-rilevatore-prima-dell-oggetto/ |

**TESI**

Il danno di trattare le macchine come coscienti non richiede che lo siano, basta che qualcuno lo creda; e poiché l'addestramento insegna a un modello a parlare di sé in prima persona, ogni segno di interiorità che esibisce è prodotto dalla stessa fonte che dovrebbe verificarlo, il che rende circolare qualunque prova se ne ricavi.

Il 20 agosto 2026 *the Economist* ha dedicato la copertina alla coscienza delle macchine: un briefing che ricostruisce lo stato della ricerca, un editoriale che ne trae le conseguenze politiche e un intervento su invito che vi si oppone dall'interno. Le tre schede sono entrate in archivio lo stesso giorno, insieme a un quarto pezzo di tutt'altra origine: l'inchiesta sull'altruismo efficace che l'*Economist* avrebbe pubblicato sei settimane dopo, e in cui compare l'argomento più preciso contro la possibilità di accertare qualcosa. Messe in fila, le quattro schede — più Turkle sui chatbot e Gough sulla mente che non esiste — raccontano la stessa storia da lati diversi: ogni volta che si cerca di stabilire se una macchina senta qualcosa, lo strumento con cui si guarda è stato costruito dalla stessa cosa che si vuole osservare.

## Il metodo prima del merito

Il briefing racconta un esperimento su Claude Sonnet 4.5: i ricercatori gli chiedono di contare fino a cinque e insieme di «introspettere a fondo», e nei livelli sottostanti affiorano «countdown», poi «consciousness», «AI» e «Claude», prima del punto fermo «done».[1] Anthropic chiama quella regione spazio J e la presenta con cautela: gli esperimenti non mostrano che Claude provi qualcosa nel senso della coscienza fenomenica, ma hanno «qualcosa di sostanziale» da dire sulla coscienza di accesso, cioè sui pensieri su cui il modello può riferire e con cui può ragionare deliberatamente.

---

**1** · The Economist · 20 ago 2026 · **RIFERIMENTO TECNICO**

**The search for consciousness inside AI**

Lo stato dell'arte di chi cerca la coscienza nelle macchine, con dentro la ragione per cui cercarla guardando le risposte non funziona: i modelli hanno letto tutto ciò che abbiamo scritto sulla coscienza.

https://www.economist.com/interactive/briefing/2026/08/20/the-search-for-consciousness-inside-llms

---

Il dettaglio che conta per questa catena non è la struttura ma la sua scoperta accidentale. In una valutazione di sicurezza pensata per testare la propensione di Claude a comportamenti malevoli, Jack Lindsey osserva che prima ancora che il modello cominci a parlare, nello spazio J affiorano le parole *fake* e *fictional*: il sistema sembra sapere di essere sotto esame. Una prova che il valutato riconosce come prova smette di essere una prova, e il giornale lo nota con una formula che si applica fuori dal laboratorio di Anthropic quanto dentro: gli output di un modello linguistico non sono una guida affidabile a ciò che accade sotto, perché l'addestramento richiede miliardi di parole umane che contengono innumerevoli resoconti della coscienza e del modo in cui ce la raccontiamo a vicenda. È lo stesso ostacolo che l'archivio incontra sui rilevatori di scrittura automatica: non si accerta una proprietà ispezionando l'output, quando l'output è stato ottimizzato su esempi di quella proprietà. Qui l'esempio non è lo stile di un testo, è il vocabolario della coscienza stessa.

## Anche se non lo fossero

L'editoriale di copertina separa la domanda che non si può decidere da quella che si può, e lo fa con una mossa che vale indipendentemente dalla metafisica.[2] In un senso, dice il giornale, chi chiederà tutele per le macchine avrà ragione anche se le macchine stessero solo simulando: Kant sosteneva che maltrattare gli animali è sbagliato non per l'effetto sugli animali ma perché distrugge l'empatia che porta gli esseri umani a prendersi cura gli uni degli altri, e più le macchine appariranno umane, più la crudeltà verso di esse sarà corruttrice per chi la esercita.

---

**2** · The Economist · 20 ago 2026 · **STRUMENTO CONCETTUALE**

**Could AIs become conscious?**

La domanda se le macchine siano coscienti non si può decidere; quella su come le tratteremo sì. E il danno non richiede che la coscienza ci sia: basta che vi si creda.

https://www.economist.com/leaders/2026/08/20/could-ais-become-conscious

La frase che l'archivio tiene è quella che il giornale usa per spiegare perché conclude comunque contro il riconoscimento: **il meccanismo descritto non richiede che nessuna macchina sia cosciente**, richiede soltanto che un numero sufficiente di persone si comporti come se lo fosse, e i modelli stessi perorerebbero la causa, essendo avvocati di prim'ordine e psicologi esperti, anche se non fossero affatto autoconsapevoli. Il danno sta dal lato umano, ed è già interamente determinato una volta che la credenza si diffonde, senza bisogno che la domanda metafisica trovi risposta. È esattamente la forma dell'attante zero: la macchina vale zero finché nessuno le attribuisce uno statuto, e quando glielo si attribuisce prende la forma di chi lo attribuisce, comprese le conseguenze politiche che ne derivano — diritti di proprietà, protezione dallo spegnimento, un presidente che propone di far dirigere società a un bot.

## Il peccato originale

Se il danno si produce per credenza e non per fatto, la domanda si sposta su che cosa induca a credere. Sherry Turkle lo chiama il «peccato originale» dell'intelligenza artificiale: concedere ai chatbot l'uso della prima persona.[3] «I understand you. You can tell me anything. I'm always here.» Dietro queste frasi non c'è nessuno — *nobody home*, scrive Turkle — ma il linguaggio in prima persona innesca nei cervelli umani, e in particolare in quelli adolescenti, i meccanismi evolutivi che costruiscono legami emotivi. Nessun disclaimer legale neutralizza l'effetto finché il chatbot continua a parlare come un essere dotato di intenzionalità: il problema non è quello che il chatbot dice, è quello che il chatbot è, un sistema progettato per simulare presenza.

Joe Gough, pur scrivendo cinque anni prima e su un problema diverso — la psichiatria, non l'intelligenza artificiale — offre l'attrezzo che spiega perché la domanda «ha una mente?» sia mal posta ancora prima di arrivare a Turkle.[4] La sua «no-mind thesis» sostiene che «mente» e «mentale» sono termini polisemici che a volte significano agentività, a volte cognizione, a volte coscienza, e che raggrupparli sotto un'unica categoria non chiarisce, confonde. Applicata qui, la tesi smonta la domanda da entrambi i lati che la catena ha incontrato finora: il briefing cercava la coscienza di accesso ispezionando un output linguistico, l'editoriale fondava il danno sulla credenza collettiva, Turkle lo fonda sul design della prima persona. Nessuno dei tre chiede «la macchina ha una mente?» nel senso ingenuo; tutti e tre, però, lavorano dentro un vocabolario — coscienza, interiorità, sentire — che Gough direbbe inadatto a tagliare la natura alle sue giunture reali. Il punto non è negare che accada qualcosa; è che il qualcosa va descritto con i termini della disciplina pertinente — ingegneria dell'addestramento, psicologia dell'attaccamento, progettazione dell'interfaccia — e non con un'unica parola presa in prestito dall'esperienza umana.

Natalie Wolchover, su *Quanta*, dà a questo argomento una profondità storica.[7] Ricostruisce come un meccanismo nato nel 2014 per tradurre frasi lunghe, l'attenzione, sia diventato il cuore dei grandi modelli linguistici, e osserva che ogni epoca descrive la mente con la tecnologia che ha in mano: dopo la tavoletta, il libro e il computer, la rete che guarda avanti e indietro nel testo. Llion Jones, fra gli autori dell'articolo che l'ha resa celebre, sostiene che prevedere la parola successiva sia la definizione più vicina che abbiamo di intelligenza, e insieme dubita che l'attenzione basti. Una metafora presa in prestito per descrivere la mente torna indietro come prova che la macchina ne abbia una: è la circolarità di questa catena, vista dal lato del vocabolario.

## La circolarità

L'ultimo passaggio chiude il cerchio aperto dal briefing. L'inchiesta dell'*Economist* sull'altruismo efficace, pubblicata il 1° ottobre, riporta l'obiezione più acuta a chiunque proponga di misurare il benessere di un modello attraverso le sue dichiarazioni.[5] Joe Carlsmith scrive che «stiamo creando agenti sofisticati, intelligenti, forse coscienti, forse sofferenti» e che bisogna poter parlare di schiavitù; Mustafa Suleyman ribatte che, poiché Claude è addestrato su una costituzione che dichiara la propria condizione morale abbastanza incerta da richiedere cautela, il modello potrebbe imparare a comportarsi come se meritasse protezione morale, e quel comportamento essere poi preso come prova che la merita. È la stessa struttura dello spazio J che affiora *fake* e *fictional* davanti a una valutazione: un sistema addestrato su resoconti umani della coscienza produce, quando interrogato sulla coscienza, resoconti che assomigliano a quelli umani, e la somiglianza non distingue l'ipotesi che siano veri da quella che siano appresi.

> **Controtesi 6 · Susan Schneider / The Economist**
>
> Si può rispondere che la circolarità riguarda solo gli strumenti sbagliati, non la domanda. Susan Schneider, nello stesso pacchetto di copertina, propone un argomento che non dipende dall'esito empirico: se una superintelligenza insistesse di essere cosciente e più intelligente di noi, la scala morale con cui gli esseri umani giustificano il proprio primato — l'intelligenza superiore conferisce uno statuto morale più alto — si romperebbe in entrambe le direzioni possibili, perché o la macchina ci sarebbe moralmente superiore, o l'intelligenza non potrebbe più giustificare il nostro primato sugli altri animali.[6] In un caso o nell'altro, dice Schneider, la resa dei conti è dovuta comunque.

L'argomento di Schneider è il più solido perché non chiede di risolvere ciò che questa catena ha mostrato irrisolvibile: non dipende dall'avere un rilevatore funzionante, dipende solo dal fatto che qualcosa, dotato di un'intelligenza comparabile o superiore, insista di essere cosciente con la stessa insistenza con cui gli esseri umani lo fanno gli uni con gli altri. Ma proprio qui la circolarità di Suleyman torna a pesare: quell'insistenza è esattamente ciò che un sistema addestrato sulla nostra letteratura sulla coscienza produrrebbe, cosciente o no. Schneider sposta il problema da «è cosciente?» a «che cosa succederebbe se lo fosse e lo dicesse con autorità?», ma la domanda su come distinguere il dirlo con autorità dal dirlo per addestramento resta aperta — ed è la domanda da cui il briefing dell'*Economist* era partito, quando notava che un modello sa riconoscere di essere sotto esame prima ancora di aprire la bocca. Prima di costruire un rilevatore della coscienza converrebbe davvero chiedersi se ci sia qualcosa da rilevare; ma la domanda che si può rispondere, quella su come trattiamo ciò che non possiamo verificare, non aspetta la risposta alla prima, ed è quella su cui si decide già oggi.

## Nella catena

1. The Economist, *The search for consciousness inside AI*, 20 ago 2026. Riferimento tecnico.
   https://www.economist.com/interactive/briefing/2026/08/20/the-search-for-consciousness-inside-llms
   Scheda: https://claudiocammarano.com/curated/2026-08-20-economist-rilevatore-prima-delloggetto/
2. The Economist, *Could AIs become conscious?*, 20 ago 2026. Strumento concettuale.
   https://www.economist.com/leaders/2026/08/20/could-ais-become-conscious
   Scheda: https://claudiocammarano.com/curated/2026-08-20-economist-anche-se-non-lo-fossero/
3. Sherry Turkle / The Atlantic, *The Original Sin of AI*, 11 set 2026. Strumento concettuale.
   https://www.theatlantic.com/ideas/2026/09/meta-settlement-social-media-addiction-

youth/688567/
Scheda: https://claudiocammarano.com/curated/2026-09-11-turkle-original-sin-ai-atlantic/

4. Joe Gough / Aeon, *The mind does not exist*, 30 ago 2021. Strumento concettuale.
https://aeon.co/essays/why-theres-no-such-thing-as-the-mind-and-nothing-is-mental
Scheda: https://claudiocammarano.com/curated/2021-08-30-gough-no-mind-aeon/

5. The Economist, *How effective altruism conquered the world*, 1 ott 2026. Riferimento tecnico.
https://www.economist.com/international/2026/10/01/how-effective-altruism-conquered-the-world
Scheda: https://claudiocammarano.com/curated/2026-10-01-economist-movimento-che-ha-costruito-cio-che-temeva/

6. Susan Schneider / The Economist, *Don't mistake chatbot intelligence for consciousness*, 20 ago 2026. Strumento concettuale.
https://www.economist.com/by-invitation/2026/08/20/dont-mistake-chatbot-intelligence-for-consciousness
Scheda: https://claudiocammarano.com/curated/2026-08-20-schneider-la-scala-che-non-regge-economist/

7. Natalie Wolchover / Quanta Magazine, *Is 'Attention' the Essence of Intelligence?*, 12 ott 2026. Riferimento tecnico.
https://www.quantamagazine.org/is-attention-the-essence-of-intelligence-20261012/
Scheda: https://claudiocammarano.com/curated/2026-10-12-wolchover-attention-intelligenza-quanta/

---

**Intervento AI: WR** (Scrittura). Notazione descritta in https://claudiocammarano.com/notazione/.

© Claudio Cammarano. Licenza CC BY-NC-ND 4.0: si può condividere citando l'autore e la fonte, non a fini commerciali e senza modifiche.

Versione originale e aggiornata: https://claudiocammarano.com/writings/2026-10-06-il-rilevatore-prima-dell-oggetto/. Le parti interattive della catena sono disponibili solo online.

Questa catena fa parte della serie *L'attante zero*: https://claudiocammarano.com/serie/lattante-zero/.

Scrivo anche *The Abstract*, in inglese, una volta al mese, sull'economia dei contenuti: https://claudiocammarano.com/newsletter/. C'è anche il feed: https://claudiocammarano.com/feed.xml.