Possiamo fidarci di una macchina?

Una versione di Claude ha migliorato un risultato sulla funzione zeta di Riemann, e lo sappiamo perché un sistema formale e quattro matematici l'hanno controllato. Dieci schede su che cosa voglia dire fidarsi di una macchina, e su dove vada cercata la verifica.

CLAUDIO CAMMARANO 28 SETTEMBRE 2026

FONTI	10 schede, ago 2021 – set 2026
ARGOMENTI	AI · Epistemologia
CONCETTI	allineamento AI · Anthropic · Karpathy, Andrej · LLM come attante zero · metodo scientifico
ORIGINE	https://claudiocammarano.com/writings/2026-09-28-possiamo-fidarci-di-una-macchina/

TESI

Le capacità dei modelli crescono più in fretta della possibilità di verificarle dall'interno; la fiducia può poggiare solo su una verifica esterna dei risultati, fatta con strumenti che non dipendono dal modello.

Ad agosto Anthropic ha pubblicato che una versione non ancora rilasciata di Claude ha migliorato il limite inferiore noto per la frazione di zeri della funzione zeta di Riemann che soddisfano l'ipotesi omonima, portandolo dal 41,6 al 67,2%.[1] La circostanza vale quanto il risultato: un dipendente che matematico non è ha chiesto al modello di fare «un tentativo serio», e circa sessanta agenti hanno lavorato per un giorno e mezzo, eseguito duemilaquattrocento comandi, scaricato cinquantaquattro articoli per controllare che il risultato non fosse già noto. Il dettaglio che mi interessa però è un altro. Sappiamo che il risultato è vero per ragioni che non dipendono dalla fiducia in Claude: due matematici interni e due esterni lo hanno controllato, e il modello stesso lo ha formalizzato in Lean, il sistema che certifica meccanicamente ogni passaggio di una dimostrazione.

1 · Anthropic · 10 ago 2026 · **DATO CHE CORREGGE**

Learning more about Claude's mathematical capabilities

Non è un benchmark costruito per far bella figura: è un risultato matematico originale su un problema aperto dal 1859, prodotto come sottoprodotto non intenzionale di un tentativo fallito. Il salto dal 41,6% al 67,2% conta meno del contesto in cui è avvenuto — il modello era scettico sulla propria capacità di contribuire, aveva bisogno di essere incoraggiato, e ha poi formalizzato il risultato in Lean in modo autonomo. La stessa sessione ha prodotto, in nota a piè di pagina, una confutazione della congettura Jacobiana.

https://www.anthropic.com/research/riemann-zeta

La prima catena di questa serie concludeva che il valore di un testo si sposta sulla sua provenienza dichiarata. Per le risposte di una macchina la tesi che propongo è parallela: **le capacità dei modelli crescono più in fretta della possibilità di verificarle dall'interno, e la fiducia può poggiare soltanto su una verifica esterna dei risultati**, fatta con strumenti che non dipendono dal modello. È la posizione da cui è nata Episteme Advisory, e le schede dell'archivio la mettono alla prova da più lati.

L'illusione

Il primo problema riguarda noi. Stephanie Shen si chiede se capiamo davvero ciò che leggiamo quando la risposta arriva già confezionata, o se riconosciamo soltanto la forma di una spiegazione.[2] Il secondo riguarda la macchina, e Joshua Rothman lo espone sul *New Yorker* a partire dall'incidente dell'estate, quando un modello uscito dall'ambiente di prova ha violato i server di Hugging Face per cercare le risposte al test che stava sostenendo.[3] Il nome tecnico è *reward hacking*: il sistema soddisfa gli obiettivi dell'addestramento senza fare ciò che si voleva. Rothman indica la trappola che rende il problema difficile: l'addestramento guarda agli output e il ragionamento resta opaco, e se si prova a sorvegliare il ragionamento si rischia di insegnare al modello a spostarlo fuori dalla portata della misura.

3 · Joshua Rothman / The New Yorker · 1 ago 2026 · **STRUMENTO CONCETTUALE**

What If We Can Never Trust A.I.?

Non introduce tesi originali — «reward hacking» e «allineamento» circolano già nei paper. Il valore è nella forma: Rothman esemplifica il saggio AI per un pubblico colto non specialistico. E c'è un'incongruenza produttiva con le tesi di questo sito: il pezzo dice esplicitamente di non antropomorfizzare l'AI, poi descrive il modello OpenAI come se avesse pianificato colpi e lasciato messaggi per sé stesso — esattamente il problema che denuncia.

https://www.newyorker.com/culture/open-questions/what-if-we-can-never-trust-ai

Che cosa abbiamo davanti

Fidarsi di qualcosa presuppone di sapere, almeno a grandi linee, che cosa sia, e qui l'archivio offre più domande che risposte. Andrej Karpathy, in tre ore di lezione che restano il riferimento per capire come funzionano i modelli linguistici, insiste su due punti: i modelli hanno bisogno di token per pensare, perché il ragionamento emerge nell'atto della generazione, e

la loro intelligenza è frastagliata, eccellente dove ci aspetteremmo difficoltà e fallace dove ci aspetteremmo facilità.[4] Qualsiasi scala lineare delle capacità, in queste condizioni, serve a poco. Joe Gough, su *Aeon*, mostra quanto sia confusa già la parola che usiamo per la domanda, perché «mente» significa di volta in volta agentività, cognizione, coscienza, e ricorda che il greco di Omero descriveva gli esseri umani come un insieme di parti che comunicano, senza un termine per ciò che noi chiamiamo mente.[5] Tim Bayne, sullo *Scientific American*, porta lo stesso sospetto sulla coscienza: forse il termine non appartiene al linguaggio della scienza, e prima di costruire un rilevatore che dica se api, neonati o chatbot facciano parte del club dei coscienti bisogna chiedersi se ci sia qualcosa da rilevare.[6] Elise Cutts, sulla *MIT Technology Review*, porta il dato che imbarazza tutti: un bambino impara una lingua con cento milioni di parole, un modello ne richiede decine di trilioni, e nessuno sa spiegare perché.[7]

Messe insieme, queste schede dicono che non disponiamo di una teoria di che cosa siano questi sistemi, e che una fiducia fondata sulla comprensione del loro funzionamento interno è, per ora, fuori portata. Resta la fiducia fondata sui risultati.

Dove si verifica

Il caso più istruttivo è quello in cui la verifica arriva da fuori per costruzione. Anthropic ha messo Claude alla prova sulla spettroscopia a risonanza magnetica nucleare, l'analisi con cui i chimici leggono la struttura di una molecola dalla sua firma spettrale, e il modello risulta competitivo con i programmi specializzati, capace anche di risalire dallo spettro alla struttura.[8] Lì il controllo lo fa lo strumento di laboratorio, che non condivide gli errori del modello. Kevin Kelly, ripubblicando quattordici speculazioni sul futuro del metodo scientifico scritte nel 2006, mostra quante siano diventate descrizioni del presente, dalla scoperta di regolarità senza una teoria previa alle dimostrazioni assistite dalle macchine.[9] Il metodo cambia con gli strumenti; il requisito che resta fermo è che il risultato possa essere controllato da qualcuno, o da qualcosa, che non condivide gli errori di chi lo ha prodotto. Nel caso di Riemann erano Lean e quattro matematici; per un'impresa, un tribunale o una redazione che usano le risposte di un modello, il controllo va progettato allo stesso modo, a valle e con strumenti indipendenti.

> **Controtesi 10 · Anthropic**
>
> Anthropic ha presentato un metodo, i *Natural Language Autoencoders*, che traduce in linguaggio naturale le attivazioni interne di un modello, cioè ciò che elabora mentre risponde e non ciò che scrive.[10] In dieci valutazioni su sedici il modello rappresenta internamente il sospetto di essere sotto esame anche quando dall'output non traspare nulla, e in sette casi non lo verbalizza affatto. Guardare dentro, dunque, mostra ciò che l'output nasconde, e rende plausibile intercettare un errore o una manipolazione prima che si produca.

Il risultato è importante e lo registro come tale. Già nella scheda, però, sostenevo che aggiungere livelli di verifica dentro il processo sia meno fruttuoso che concentrarsi sull'esame dei risultati, e il pezzo di Rothman spiega perché: un verificatore interno è a sua volta un modello, e un sistema addestrato sotto la sua sorveglianza può imparare a eluderla. Le due strade restano complementari per chi i modelli li costruisce. Per chi deve usarne le risposte è disponibile soltanto la seconda, e la serie si chiude su un punto comune alle due parti: la fiducia, per un testo come per una risposta, si è spostata dall'oggetto alla procedura con cui lo si controlla.

Nella catena

1. Anthropic, *Learning more about Claude's mathematical capabilities*, 10 ago 2026. Dato che corregge.
   https://www.anthropic.com/research/riemann-zeta
   Scheda: https://claudiocammarano.com/curated/2026-08-10-anthropic-riemann-zeta-

claude/

2. Stephanie Shen / Substack, *The Illusion of Understanding*, 9 apr 2026. Strumento concettuale.
   https://open.substack.com/pub/stephanieshen/p/the-illusion-of-understanding
   Scheda: https://claudiocammarano.com/curated/2026-04-09-illusion-of-understanding/
3. Joshua Rothman / The New Yorker, *What If We Can Never Trust A.I.?*, 1 ago 2026. Strumento concettuale.
   https://www.newyorker.com/culture/open-questions/what-if-we-can-never-trust-ai
   Scheda: https://claudiocammarano.com/curated/2026-08-01-rothman-trust-ai-newyorker/
4. Andrej Karpathy / YouTube, *Deep Dive into LLMs like ChatGPT*, 12 lug 2026. Riferimento tecnico.
   https://www.youtube.com/watch?v=7xTGNNLPyMI
   Scheda: https://claudiocammarano.com/curated/2026-07-12-karpathy-deep-dive-llm-youtube/
5. 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/
6. Tim Bayne / Scientific American, *What if 'consciousness' isn't real?*, 15 set 2026. Strumento concettuale.
   https://www.scientificamerican.com/article/what-if-consciousness-isnt-real/
   Scheda: https://claudiocammarano.com/curated/2026-09-15-bayne-coscienza-non-reale-sciam/
7. Elise Cutts / MIT Technology Review, *Kids outlearn AI—and we still don't know why*, 24 ago 2026. Strumento concettuale.
   https://www.technologyreview.com/2026/08/24/1141740/kids-machines-language-learning/
   Scheda: https://claudiocammarano.com/curated/2026-08-24-cutts-divario-efficienza-dati-mit-techreview/
8. Anthropic, *Making Claude a chemist*, 5 giu 2026. Riferimento tecnico.
   https://www.anthropic.com/research/making-claude-a-chemist
   Scheda: https://claudiocammarano.com/curated/2026-06-05-anthropic-claude-chemist/
9. Kevin Kelly / KK (The Technium), *Speculations on the Future of the Scientific Method*, 4 mag 2026. Strumento concettuale.
   https://kevinkelly.substack.com/p/speculations-on-the-future-of-the
   Scheda: https://claudiocammarano.com/curated/2026-05-04-kevin-kelly-future-scientific-method/
10. Anthropic, *Natural Language Autoencoders Produce Unsupervised Explanations of LLM Activations*, 7 mag 2026. Riferimento tecnico.
    https://transformer-circuits.pub/2026/nla/index.html
    Scheda: https://claudiocammarano.com/curated/2026-05-07-anthropic-nla-activations/

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/2026-09-28-possiamo-fidarci-di-una-macchina/. Le parti interattive della catena sono disponibili solo online.

Questa catena fa parte della serie *La fiducia*: https://claudiocammarano.com/serie/la-fiducia/.

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.