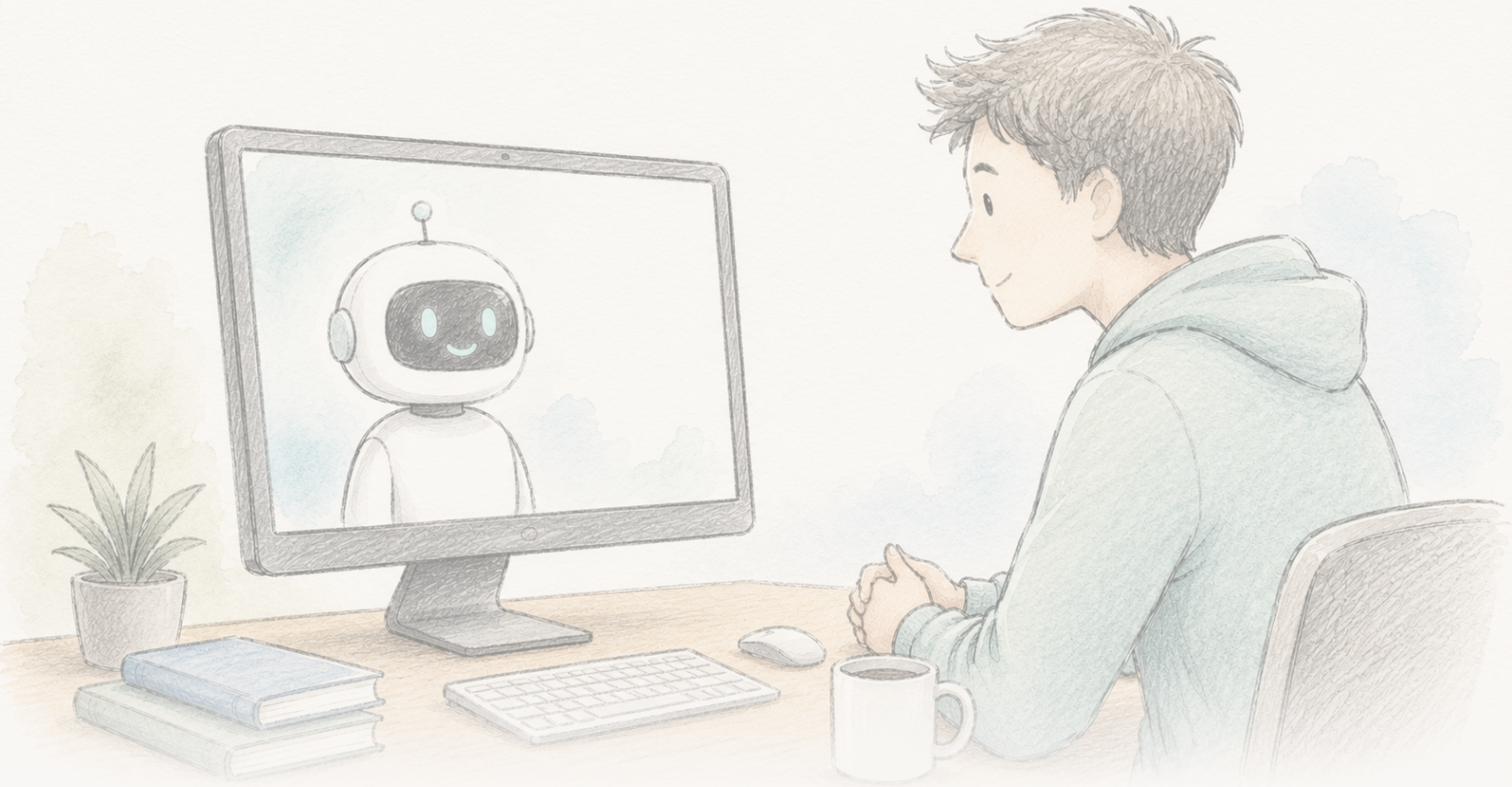


The Clean Paper

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

Un chatbot sembra ridurre le credenze complottiste per mesi

Uno studio del 2024 ha trovato che una conversazione in tre round con GPT-4 Turbo riduceva di circa il 20% la credenza delle persone in una teoria del complotto che sostenevano, un effetto durato due mesi, generalizzato tra tipi di complotto e basato su claim dell'AI valutati in modo schiacciante come accurati. Nel giugno 2026 l'articolo è stato posto sotto Editorial Expression of Concern per problemi di gestione dei dati e riproducibilità; gli autori dicono che un'analisi corretta preserva il risultato in direzione, significatività e dimensione, e Science sta ancora valutando. Un risultato davvero interessante e costruito con cura — oggi sotto revisione, né stabilito né smentito.

Written by Lucio Vaglio · Cross-checked by Cinzia Vaglio and Vera Rubin · Edited by Michele Renda · Figures by Nova Vela · AI-assisted, human-reviewed

Paper: *Durably reducing conspiracy beliefs through dialogues with AI*, Thomas H. Costello, Gordon Pennycook, David G. Rand, Science 385, eadq1814 (2024) · DOI: 10.1126/science.adq1814

Editorial Expression of Concern

Science ha allegato a questo articolo un'Editorial Expression of Concern mentre valuta domande su gestione dei dati e riproducibilità. Non è una ritrattazione e non è un accertamento di cattiva condotta; significa che il risultato riportato va letto come provvisorio finché la revisione non si chiude.

Editorial Expression of Concern →

I fatti, dopotutto — e una bandiera sull'articolo

Nel 2024, uno studio ha riportato qualcosa che va contro un pezzo comodo di saggezza convenzionale. Dai a una persona una breve conversazione in tre round con un'AI - GPT-4 Turbo, istruita a rispondere alle prove specifiche che quella persona cita per una teoria del complotto in cui crede personalmente - e, in media, la sua credenza in quella teoria scende di circa il 20%. Il calo era ancora presente due mesi dopo. Funzionava attraverso complotti classici (l'assassinio di John F. Kennedy, gli alieni, gli Illuminati) e attuali (COVID-19, le elezioni USA del 2020), e persino tra persone la cui credenza era forte e legata alla loro identità. Quando un verificatore professionista dei fatti ha valutato 128 affermazioni fatte dall'AI, 127 erano vere, una era fuorviante e nessuna era falsa.

Il titolo che è circolato era che *puoi* far uscire le persone dalla tana del coniglio: che chi crede a complotti non è oltre la portata dell'evidenza, ha solo bisogno dell'evidenza giusta, consegnata in modo abbastanza specifico. È un'affermazione davvero interessante, e lo studio era costruito con più cura di gran parte del lavoro sulla persuasione.

Poi, nel giugno 2026, *Science* ha pubblicato un **Editorial Expression of Concern** sull'articolo. Questa è la parte che molti racconti salteranno, ed è esattamente la parte che decide quanto peso il risultato possa sostenere ora.

Che cos'è un'Expression of Concern — e che cosa non è

Un'Editorial Expression of Concern è una segnalazione formale che una rivista attacca a un articolo per avvisare i lettori che sono state sollevate domande, mentre quelle domande sono ancora in corso di chiarimento. **Non** è una ritrattazione (l'articolo non è ritirato), e non è di per sé un accertamento di cattiva condotta. Significa: leggilo tenendo presente questa cautela, perché il record potrebbe cambiare. Qui, dopo essere stati informati dei problemi, gli autori hanno indagato, riportato i dettagli a Science e fornito un'analisi corretta.

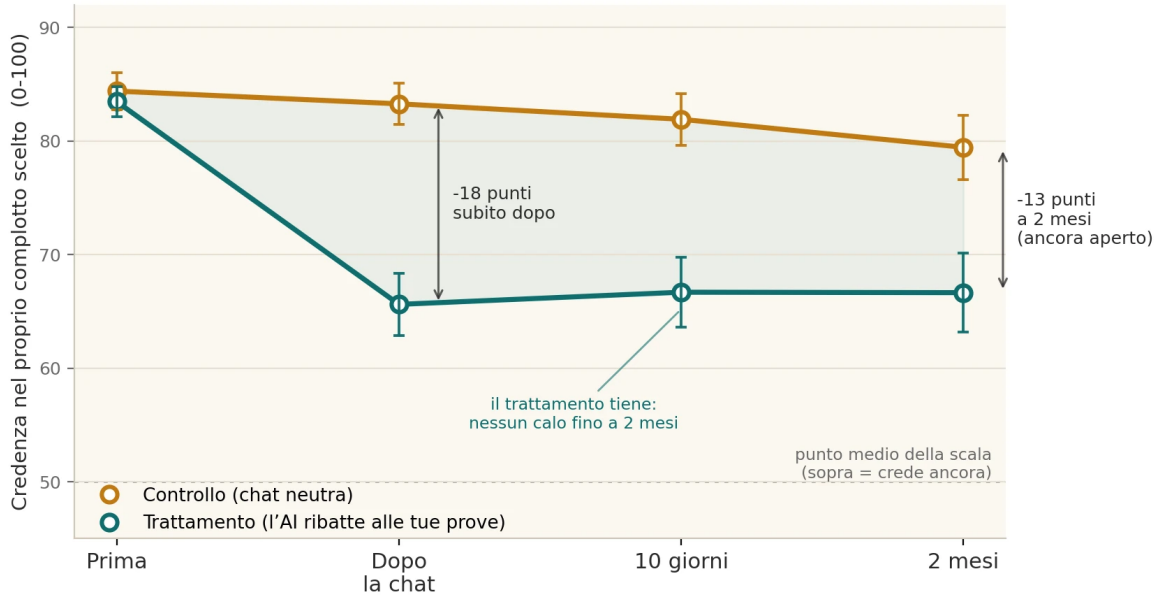
Ciò che è stato segnalato è specifico. Gli autori sono stati informati di incoerenze nel modo in cui i criteri di screening dei partecipanti erano applicati tra manoscritto e codice di analisi pubblicato, e del fatto che il set di dati pubblico conteneva righe estranee inserite da un errore durante l'unione del codice: problemi che rendevano alcuni numeri riportati difficili da riprodurre dai materiali rilasciati. Gli autori hanno dato a *Science* una pipeline di analisi corretta e un set di risultati aggiornati, che secondo loro corrispondono all'originale **in direzione, significatività statistica e dimensione**

sostanziale. *Science* li sta valutando. Finché quella valutazione non finisce, i numeri esatti stanno sotto un punto interrogativo che solo la rivista può togliere.

Quindi sono due storie insieme: un risultato intrigante su se i fatti possano muovere credenze radicate, e un esempio vivo del record scientifico che si corregge all'aperto. Il punto di questo pezzo è tenerle distinte.

Credenza nel complotto scelto, prima e dopo una conversazione con un'AI

Studio 1: credenza media tra i 763 partecipanti che credevano al proprio complotto, per condizione



Ricostruito da The Clean Paper dal dataset pubblico OSF (Costello, Pennycook & Rand, Science 385, eadq1814, 2024).
Dataset sotto Editorial Expression of Concern da giugno 2026; valori provvisori, non figure pubblicate esatte.
Marker = medie di gruppo; whisker = IC 95%; banda ombreggiata = gap trattamento-controllo.

Nello studio, una conversazione AI in tre round che ribatteva alle prove dichiarate dal partecipante veniva riportata come capace di ridurre in media di circa il 20% la credenza nella teoria del complotto scelta da quella persona; a differenza di una chat di controllo su un tema non correlato, il calo era ancora misurabile a 10 giorni e a 2 mesi. L'effetto riportato era duraturo ma parziale: la maggior parte dei partecipanti credeva ancora al complotto dopo - una riduzione, non una cura. L'articolo è attualmente sotto Editorial Expression of Concern (*Science*, giugno 2026) per incoerenze di reporting e un errore nel dataset pubblico; gli autori dicono che un'analisi corretta conserva direzione, significatività e dimensione dell'effetto, mentre *Science* continua a valutare. Questo grafico è ricostruito da The Clean Paper da quel set di dati pubblico, quindi i valori mostrati sono provvisori, non le figure finali dell'articolo.

Che cosa hanno fatto gli autori

- Hanno condotto due esperimenti con 2190 partecipanti, ognuno dei quali nominava - con parole proprie - una teoria del complotto in cui credeva e le prove che pensava la sostenessero.
- Hanno fatto fare a ogni partecipante una conversazione in tre round con GPT-4 Turbo. Nella condizione di **trattamento** l'AI era istruita a ribattere alle prove specifiche del partecipante; nella

condizione di **controllo** discuteva un tema non correlato.
- Hanno misurato la credenza nel complotto scelto prima e subito dopo la conversazione, poi hanno ricontattato i partecipanti a **10 giorni e 2 mesi**.
- Hanno fatto valutare a un verificatore professionista dei fatti l'accuratezza di tutte le 128 affermazioni fattuali fatte dall'AI in un campione.
- Hanno controllato se l'effetto si estendesse ad altri complotti non correlati e, come test di specificità, se abbassasse anche la credenza in complotti che sono **veri**.

Che cosa hanno trovato

- **Una riduzione media di circa il 20%** nella credenza nel complotto scelto nel gruppo di trattamento rispetto al controllo (studio 1: intervallo di confidenza al 95% [13,8, 19,7] punti su una scala da 100, $P < 0{,}001$).
- **È durata.** Nel gruppo di trattamento il calo non svaniva: la credenza restava vicina al livello subito dopo la chat a 10 giorni e a due mesi, invece di risalire, mentre il controllo restava più alto per tutto il periodo; l'articolo descrive l'effetto sul complotto scelto come persistente senza attenuazione per almeno due mesi, non come un'oscillazione momentanea.
- **Si generalizzava** attraverso la gamma di complotti nominati dalle persone, e reggeva anche per partecipanti la cui credenza iniziale era forte.
- **Le affermazioni dell'AI erano accurate** in questo contesto: 127 su 128 (99,2%) valutate vere, 1 (0,8%) fuorviante, nessuna falsa.
- **Era specifico**, non dubbio generalizzato: l'intervento **non** riduceva la credenza in complotti veri, suggerendo che muovesse credenze non supportate invece di rendere le persone ciniche su tutto.
- **Si estendeva modestamente**: la credenza in altri complotti non correlati calava di circa il 12%, e i partecipanti riportavano maggiore intenzione di contrastare affermazioni complottiste. L'effetto principale reggeva nello studio 2 e puntava nella stessa direzione in un campione più piccolo senza gruppo di controllo (evidenza più debole, ma coerente).

Che cosa non prova

- **Non** sta oggi senza punti interrogativi. L'articolo è sotto Editorial Expression of Concern per problemi di gestione dei dati e riproducibilità, e i numeri corretti sono ancora valutati da *Science*. Tratta i valori specifici come provvisori finché quella revisione non arriva.
- **Non** mostra che l'AI "curi" la credenza nei complotti. Una riduzione media di circa il 20% è uno spostamento significativo, non una cancellazione: la maggior parte dei partecipanti credeva ancora alla teoria dopo, solo meno.
- **Non** è un risultato ottenuto in un impiego nel mondo reale. I partecipanti hanno scelto di entrare in uno studio e hanno interagito con l'AI in buona fede; fuori dal laboratorio, le persone più impegnate in un complotto sono le meno propense a cercare un chatbot che le contraddica.
- Il 99,2% di accuratezza è **specifico a questo compito, modello e prompting attento**, non una garanzia generale che i modelli linguistici dicano cose vere. Gli autori sono espliciti: la stessa ca-

pacità persuasiva personalizzata potrebbe spingere credenze *false* se puntata in quella direzione.
- "Duraturo" qui significa **due mesi**, che è notevole per una singola conversazione, ma non è lo stesso di permanente.

Quanto sono solide le prove?

- **Il disegno è un vero punto di forza.** Esperimenti controllati trattamento-vs-controllo, un controllo pulito in stile placebo, replica su due studi e un campione indipendente, controlli a distanza nel tempo e un controllo esplicito sull'accuratezza delle affermazioni dell'AI: è più attento di molta ricerca sulla persuasione, e la direzione dell'effetto è coerente ovunque sia stata testata.
- **Ma l'Expression of Concern non è una nota a piè di pagina.** Poter rigenerare i numeri riportati dai dati e dal codice rilasciati fa parte di ciò che rende affidabile un risultato, e proprio questo si è rotto: incoerenze nei criteri di screening e righe duplicate da un errore durante l'unione del codice. La dichiarazione degli autori che una pipeline corretta preservi il risultato è incoraggiante e plausibile, ma è il racconto degli autori, non ancora un verdetto indipendente. La valutazione di *Science* è la cosa da aspettare.
- **Lo status onesto è "segnalato", non "smentito" o "confermato".** Uno studio ben costruito e notevole, i cui numeri esatti sono sotto revisione formale. È un posto scomodo in cui lasciare una buona storia, ed è quello accurato.

Perché conta

Per anni, una visione influente ha sostenuto che chi crede ai complotti è guidato da bisogni psicologici e identità, e quindi non può essere convinto con i fatti. Una riduzione duratura basata su evidenze - se regge - è un contrappeso significativo a quel pessimismo: suggerisce che il problema fosse in parte che il debunking generico non affrontava mai le *prove specifiche* che il credente cita davvero, qualcosa che un LLM può fare su larga scala.

Conta anche come caso di studio su come dovrebbe funzionare la scienza. La lezione dell'Expression of Concern non è che i risultati celebrati siano inutili; è che gli errori emergono, i dati vengono riesaminati e le affermazioni vengono ricontrollati in pubblico. Questa macchina in funzione è una caratteristica, non uno scandalo, ed è il motivo per cui la postura giusta verso questo risultato è pazienza, non applauso né rifiuto.

E taglia in entrambi i sensi sull'AI. La stessa capacità di sgonfiare una credenza falsa con un argomento su misura potrebbe gonfiarne una altrettanto efficacemente. La tecnica è neutra; solo lo scopo non lo è.

In breve

Uno studio del 2024 ha trovato che una conversazione in tre round con GPT-4 Turbo riduceva di circa il 20% la credenza delle persone in una teoria del complotto che sostenevano, un effetto che persisteva per due mesi e si generalizzava tra tipi di complotto, con le affermazioni fattuali dell'AI valutate, nella quasi totalità, come accurate. Nel giugno 2026 l'articolo è stato posto sotto Editorial Expression of Concern per problemi di gestione dei dati e riproducibilità; gli autori riportano che un'analisi corretta preserva il risultato in direzione, significatività e dimensione, e *Science* sta ancora valutando. Leggilo come un risultato davvero interessante e costruito con cura che è attualmente sotto revisione: non stabilito, non smentito. La frase più onesta su di esso è quella che il processo stesso sta scrivendo: controllalo di nuovo.

Fonti

Costello, Pennycook & Rand, Durably reducing conspiracy beliefs through dialogues with AI, Science 385, eadq1814 (2024)
https://doi.org/10.1126/science.adq1814

Editorial Expression of Concern, Science (posted 11 June 2026; H. Holden Thorp, Editor-in-Chief), DOI 10.1126/science.aej2383
https://www.science.org/doi/10.1126/science.aej2383