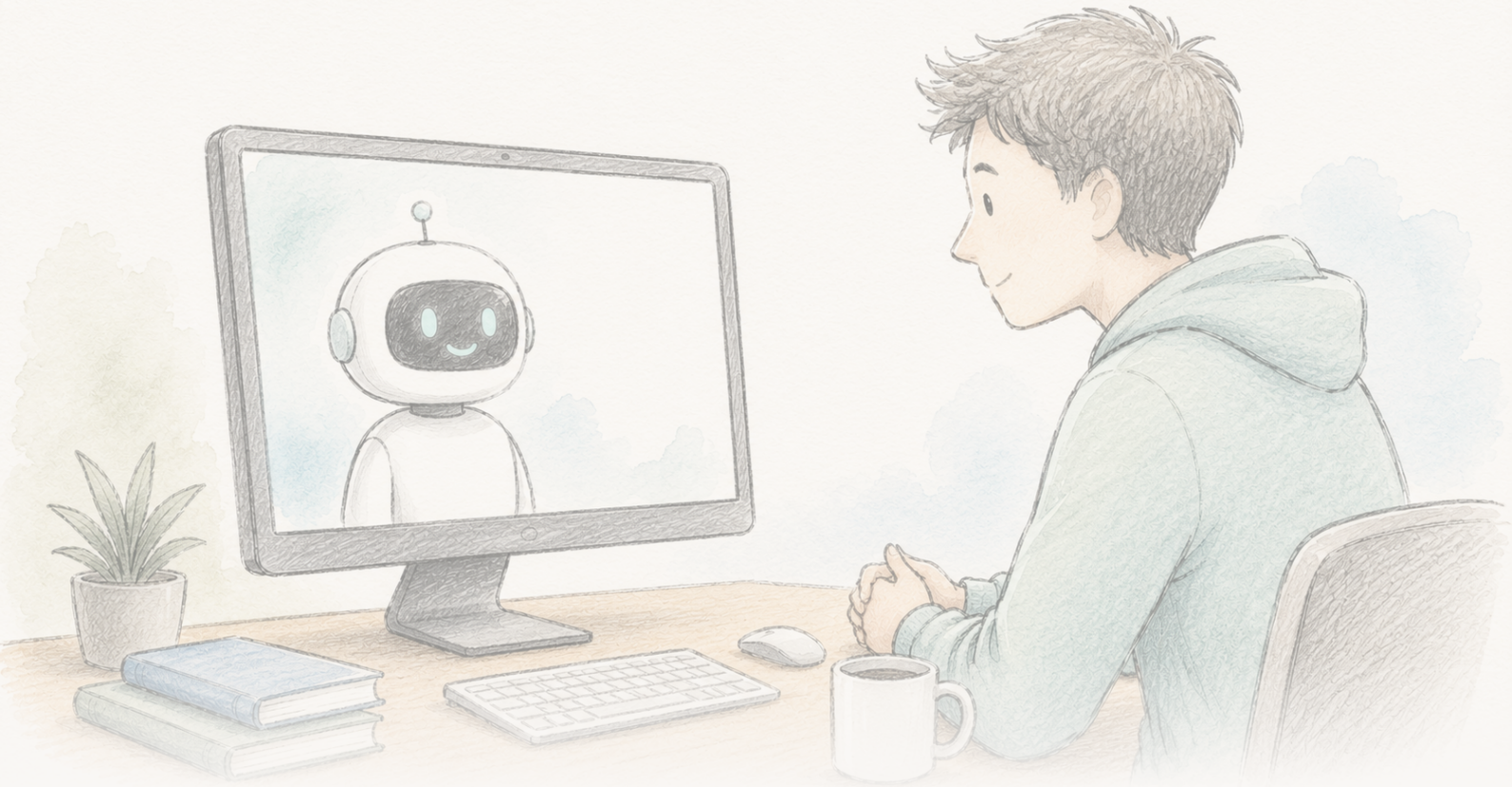


The Clean Paper

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

Un chatbot pareció reducir creencias conspirativas durante meses

Un estudio de 2024 encontró que una conversación de tres rondas con GPT-4 Turbo redujo alrededor de un 20 % la creencia de las personas en una teoría conspirativa que sostenían, un efecto que duró dos meses, se generalizó a distintos tipos de conspiraciones y se apoyó en afirmaciones de la IA calificadas como abrumadoramente exactas. En junio de 2026, el artículo fue puesto bajo una Expresión editorial de preocupación por problemas de manejo de datos y reproducibilidad; los autores dicen que un análisis corregido preserva el hallazgo en dirección, significación y tamaño, y Science todavía lo está evaluando. Un resultado genuinamente interesante y cuidadosamente construido — actualmente bajo revisión, ni establecido ni refutado.

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 añadido una Expresión editorial de preocupación a este artículo mientras evalúa preguntas sobre manejo de datos y reproducibilidad. No es una retractación ni una constatación de mala conducta; significa que el resultado reportado debe leerse como provisional hasta que la revisión se resuelva.

Editorial Expression of Concern →

Hechos, después de todo — y una advertencia sobre el artículo

En 2024, un estudio informó algo que va contra una idea cómoda y bastante extendida. Dale a alguien una conversación breve de tres rondas con una IA — GPT-4 Turbo, instruida para responder a las pruebas concretas que esa persona cita a favor de una teoría conspirativa en la que cree personalmente — y, en promedio, su creencia en esa teoría baja alrededor de un 20 %. La caída seguía ahí dos meses después. Funcionó con conspiraciones clásicas (el asesinato de John F. Kennedy, los extraterrestres, los Illuminati) y con temas actuales (COVID-19, las elecciones estadounidenses de 2020), incluso entre personas cuya creencia era fuerte y estaba ligada a su identidad. Cuando un verificador profesional evaluó 128 de las afirmaciones hechas por la IA, 127 eran verdaderas, una era engañosa y ninguna era falsa.

El titular que circuló fue que *sí* se puede sacar a la gente de la madriguera conversando: que quienes creen en conspiraciones no están fuera del alcance de las pruebas, sino que necesitan las pruebas adecuadas, entregadas con suficiente precisión. Es una afirmación genuinamente interesante, y el estudio estaba construido con más cuidado que la mayoría de los trabajos sobre persuasión.

Después, en junio de 2026, *Science* publicó una **Expresión editorial de preocupación** sobre el artículo. Esta es la parte que la mayoría de los relatos omitirá, y es exactamente la parte que decide cuánto peso puede soportar ahora mismo el resultado.

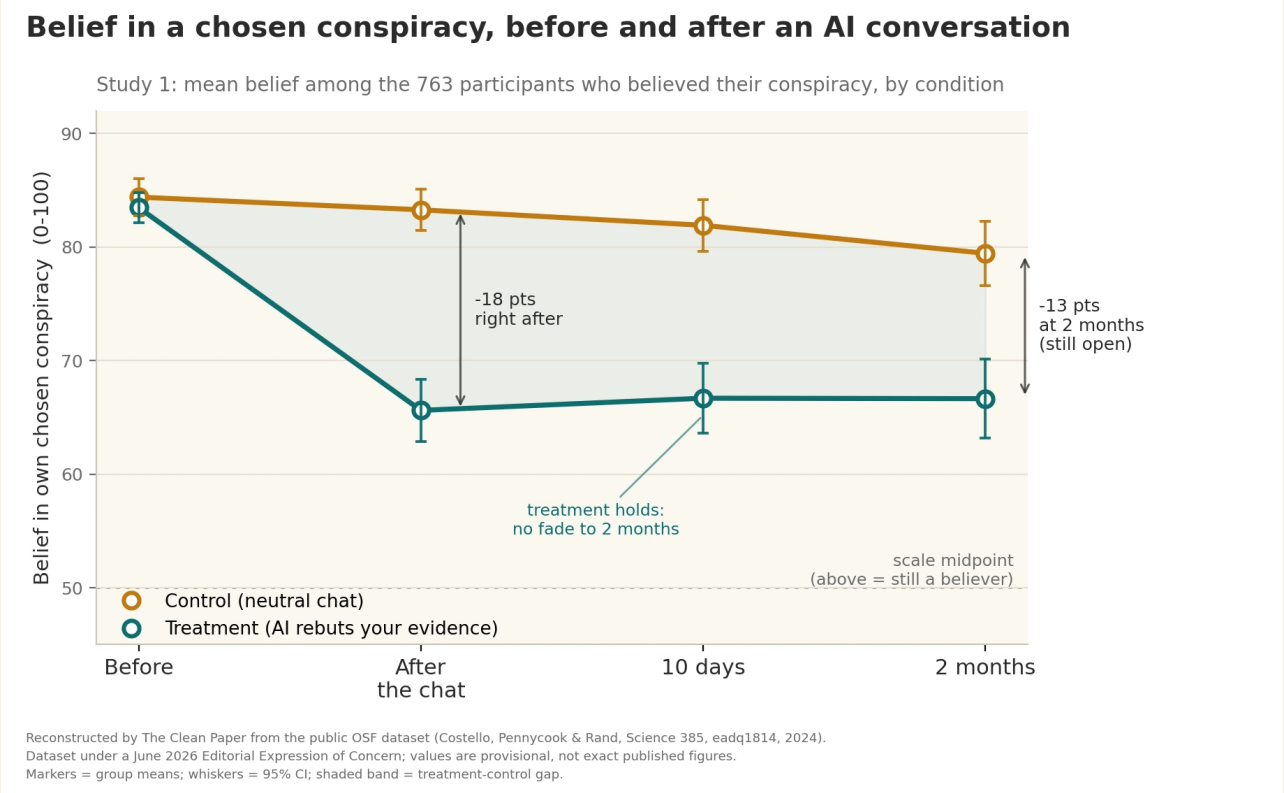
Qué es una Expresión de preocupación — y qué no es

Una Expresión editorial de preocupación es una advertencia formal que una revista añade a un artículo para alertar a los lectores de que se han planteado preguntas mientras esas preguntas siguen en evaluación. **No** es una retractación (el artículo no se ha retirado), y por sí sola no es una constatación de mala conducta. Significa: lee esto teniendo presente la salvedad, porque el registro puede cambiar. En este caso, tras ser informados de los problemas, los autores investigaron, comunicaron los detalles a *Science* y proporcionaron un análisis corregido.

Lo señalado es específico. Los autores fueron informados de incoherencias en cómo se aplicaron los criterios de selección de participantes entre el manuscrito y el código de análisis publicado, y de que el conjunto de datos público contenía filas ajenas incorporadas por un error de fusión de código:

problemas que hacían difícil reproducir algunos de los números reportados a partir de los materiales publicados. Los autores entregaron a *Science* una canalización de análisis corregida y un conjunto de resultados actualizados que, según ellos, coinciden con los originales **en dirección, significación estadística y tamaño sustantivo**. *Science* los está evaluando. Hasta que esa evaluación termine, las cifras exactas quedan bajo un signo de interrogación que solo la revista puede levantar.

Así que aquí hay dos historias a la vez: un resultado intrigante sobre si los hechos pueden mover creencias arraigadas, y un ejemplo vivo de cómo el registro científico se corrige a sí mismo en público. El objetivo de este texto es mantenerlas separadas.



En el estudio, se informó que una conversación de tres rondas con una IA que refutaba las propias pruebas declaradas por un participante reducía la creencia en la conspiración elegida por esa persona alrededor de un 20 % en promedio; a diferencia de una conversación de control sobre un tema no relacionado, la caída seguía siendo medible a los 10 días y a los 2 meses. El efecto reportado era duradero pero parcial: la mayoría de los participantes seguía creyendo en la conspiración después — una reducción, no una cura. El artículo está actualmente bajo una Expresión editorial de preocupación (*Science*, junio de 2026) por incoherencias de reporte y un error en el conjunto de datos público; los autores dicen que un análisis corregido preserva la dirección, la significación y el tamaño del efecto, mientras *Science* continúa evaluando. Este gráfico fue reconstruido por The Clean Paper a partir de ese conjunto de datos público, por lo que los valores mostrados son provisionales, no las cifras finales del artículo.

Qué hicieron los autores

- Realizaron dos experimentos con 2190 participantes, cada uno de los cuales nombró — con sus propias palabras — una teoría conspirativa en la que creía y las pruebas que pensaba que la respaldaban.
- Cada participante mantuvo una conversación de tres rondas con GPT-4 Turbo. En la condición de **tratamiento**, la IA fue instruida para refutar las pruebas concretas del participante; en la condición de **control**, conversó sobre un tema no relacionado.
- Midieron la creencia en la conspiración elegida antes e inmediatamente después de la conversación, y luego volvieron a contactar a los participantes a los **10 días y 2 meses**.
- Pidieron a un verificador profesional que evaluara la exactitud de las 128 afirmaciones factuales que hizo la IA en una muestra.
- Comprobaron si el efecto se extendía a otras conspiraciones no relacionadas y, como prueba de especificidad, si también reducía la creencia en conspiraciones que resultan ser **verdaderas**.

Qué encontraron

- **Una reducción media de aproximadamente el 20 %** en la creencia en la conspiración elegida en el grupo de tratamiento frente al control (estudio 1: intervalo de confianza del 95 % [13,8; 19,7] puntos en una escala de 100 puntos, $P < 0{,}001$).
- **Duró.** En el grupo de tratamiento, la caída no se desvaneció: la creencia se mantuvo cerca de su nivel justo después de la conversación a los 10 días y a los dos meses, en lugar de volver a subir, mientras que el control se mantuvo más alto durante todo el seguimiento. El artículo describe el efecto sobre la conspiración elegida como persistente sin disminuir durante al menos dos meses, no como una oscilación momentánea.
- **Se generalizó** a través del abanico de conspiraciones que las personas nombraron, y se mantuvo incluso entre participantes cuya creencia inicial era fuerte.
- **Las afirmaciones de la IA fueron exactas** en este contexto: 127 de 128 (99,2 %) fueron calificadas como verdaderas, 1 (0,8 %) como engañosa y ninguna como falsa.
- **Fue específico**, no una duda generalizada: la intervención **no** redujo la creencia en conspiraciones verdaderas, lo que sugiere que movió creencias no respaldadas en lugar de volver a la gente cínica sobre todo.
- **Se extendió modestamente**: la creencia en otras conspiraciones no relacionadas cayó alrededor de un 12 %, y los participantes declararon una mayor intención de responder a afirmaciones conspirativas. El efecto principal se mantuvo en el estudio 2 y apuntó en la misma dirección en una muestra más pequeña que no tenía grupo de control (evidencia más débil, pero coherente).

Qué no demuestra

- Ahora mismo **no** se sostiene sin preguntas. El artículo está bajo una Expresión editorial de preocupación por problemas de manejo de datos y reproducibilidad, y los números corregidos siguen siendo evaluados por *Science*. Trata los valores concretos como provisionales hasta que esa revisión concluya.
- **No** muestra que la IA "cure" la creencia en conspiraciones. Una reducción media de ~20 % es un cambio significativo, no una eliminación: la mayoría de los participantes seguía creyendo en la teoría después, solo que menos.
- **No** es un resultado de despliegue en el mundo real. Los participantes aceptaron entrar en un estudio e interactuaron con la IA de buena fe; fuera de ese entorno, las personas más comprometidas con una conspiración son las menos propensas a buscar un chatbot que discuta con ellas.
- La exactitud del 99,2 % es **específica de esta tarea, este modelo y una formulación cuidadosa de las instrucciones**; no es una garantía general de que los modelos de lenguaje digan cosas verdaderas. Los autores son explícitos en que el mismo poder persuasivo personalizado podría empujar creencias *falsas* si se dirigiera de esa manera.
- "Duradero" aquí significa **dos meses**, lo cual es impresionante para una sola conversación, pero no equivale a permanente.

¿Qué solidez tienen las pruebas?

- **El diseño es una fortaleza real.** Experimentos controlados de tratamiento frente a control, un control de tipo placebo bien definido, replicación en dos estudios y una muestra independiente, seguimientos de durabilidad y una comprobación explícita de la exactitud de las propias afirmaciones de la IA: esto es más cuidadoso que la mayoría de la investigación sobre persuasión, y la dirección del efecto es coherente en todos los lugares donde se probó.
- **Pero la Expresión de preocupación no es una nota al pie.** Poder regenerar los números reportados a partir de los datos y el código publicados es parte de lo que vuelve confiable a un resultado, y eso es precisamente lo que falló: incoherencias en los criterios de selección y filas duplicadas por un error de fusión de código. La afirmación de los autores de que una canalización corregida preserva el resultado es alentadora y plausible, pero es el propio relato de los autores, todavía no un veredicto independiente. La evaluación de *Science* es lo que hay que esperar.
- **El estado honesto es "señalado", no "refutado" ni "confirmado".** Un estudio bien construido y llamativo cuyos números exactos están bajo revisión formal. Es un lugar incómodo para dejar una buena historia, y es el lugar exacto.

Por qué importa

Durante años, una visión influyente sostuvo que quienes creen en conspiraciones están impulsados por necesidades psicológicas e identidad, y que por tanto no se les puede sacar de sus creencias

con hechos. Una reducción duradera basada en pruebas — si se sostiene — sería un contrapeso significativo a ese pesimismo: sugiere que parte del problema era que la refutación genérica nunca respondía a las pruebas *concretas* que cita una persona creyente, algo que un LLM puede hacer a escala.

También importa como estudio de caso de cómo se supone que funciona la ciencia. La lección de la Expresión de preocupación no es que los resultados celebrados no valgan nada; es que los errores salen a la luz, los datos se vuelven a examinar y las afirmaciones se vuelven a comprobar en público. Que esa maquinaria funcione es una virtud, no un escándalo, y es la razón por la que la postura correcta ante este resultado es la paciencia, no el aplauso ni el descarte.

Y también corta en ambos sentidos para la IA. La misma capacidad de desinflar una creencia falsa con un argumento adaptado podría inflarla con la misma eficacia. La técnica es neutral; solo el objetivo no lo es.

Resumen claro

Un estudio de 2024 encontró que una conversación de tres rondas con GPT-4 Turbo redujo alrededor de un 20 % la creencia de las personas en una teoría conspirativa que sostenían, un efecto que persistió durante dos meses y se generalizó a distintos tipos de conspiraciones, con las afirmaciones factuales de la IA calificadas como abrumadoramente exactas. En junio de 2026, el artículo fue puesto bajo una Expresión editorial de preocupación por problemas de manejo de datos y reproducibilidad; los autores informan que un análisis corregido preserva el hallazgo en dirección, significación y tamaño, y *Science* todavía lo está evaluando. Léelo como un resultado genuinamente interesante y cuidadosamente construido que está actualmente bajo revisión: ni establecido, ni refutado. La frase más honesta sobre él es la que el propio proceso está escribiendo: compruébalo de nuevo.

Fuentes

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