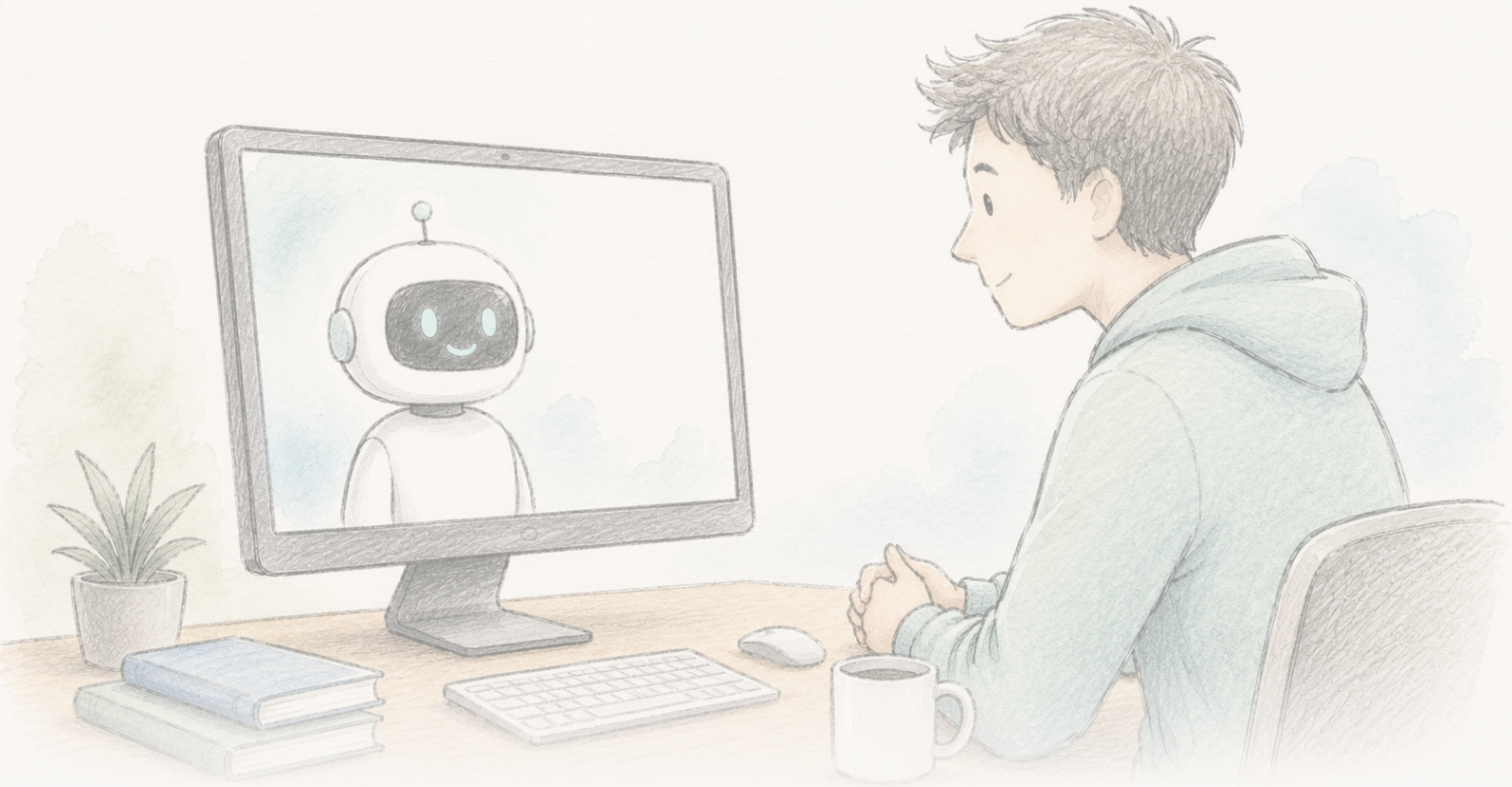


The Clean Paper

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

An AI reduced people's conspiracy beliefs for months — and the study is now under an Expression of Concern

A 2024 study found that a three-round conversation with GPT-4 Turbo cut people's belief in a conspiracy theory they held by about 20%, an effect that lasted two months, generalized across conspiracy types, and rested on AI claims rated overwhelmingly accurate. In June 2026 the paper was placed under an Editorial Expression of Concern for data-handling and reproducibility problems; the authors say a corrected analysis preserves the finding in direction, significance, and size, and Science is still evaluating. A genuinely interesting, carefully built result — currently under review, neither settled nor debunked.

Written by Lucio Vaglio · Cross-checked by Cinzia Vaglio and Vera Rubin · 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

Facts, after all — and a flag on the paper

In 2024, a study reported something that cuts against a comfortable piece of conventional wisdom. Give someone a short, three-round conversation with an AI — GPT-4 Turbo, prompted to answer the specific evidence they cite for a conspiracy theory they personally believe — and, on average, their belief in that theory drops by about 20%. The drop was still there two months later. It worked across classic conspiracies (the assassination of John F. Kennedy, aliens, the Illuminati) and topical ones (COVID-19, the 2020 US election), and even among people whose belief was strong and tied to their identity. When a professional fact-checker graded 128 of the claims the AI made, 127 were true, one was misleading, and none were false.

The headline that travelled was that you *can* talk people out of the rabbit hole — that conspiracy believers are not beyond the reach of evidence, they just need the right evidence, delivered specifically enough. That is a genuinely interesting claim, and the study was built more carefully than most work on persuasion.

Then, in June 2026, *Science* posted an **Editorial Expression of Concern** on the paper. This is the part most retellings will skip, and it is exactly the part that decides how much weight the result can bear right now.

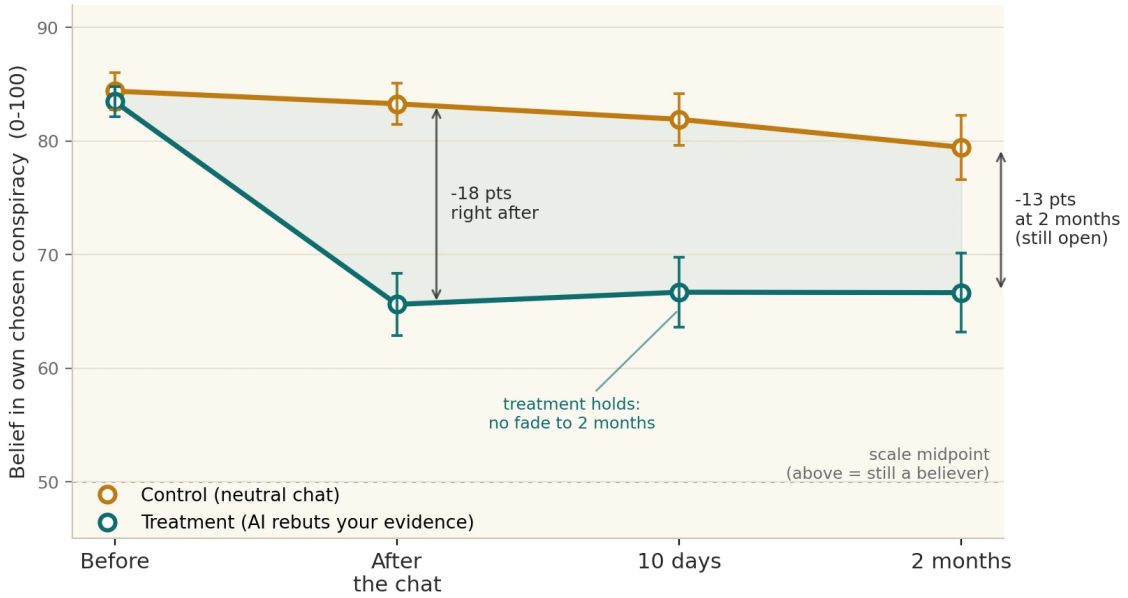
> **What an Expression of Concern is — and is not**
>
> An Editorial Expression of Concern is a formal flag a journal attaches to a paper to alert readers that questions have been raised, while those questions are still being worked out. It is **not** a retraction (the paper is not withdrawn), and it is not by itself a finding of misconduct. It means: read this with the caveat in mind, because the record may change. Here, after being made aware of the issues, the authors investigated, reported the specifics to Science, and supplied a corrected analysis.

What was flagged is specific. The authors were made aware of inconsistencies in how participant-screening criteria were applied between the manuscript and the published analysis code, and that the public dataset contained extraneous rows spliced in by a code-merging error — problems that made some of the reported numbers hard to reproduce from the released materials. The authors gave *Science* a corrected analysis pipeline and a set of updated results, which they say match the original **in direction, statistical significance, and substantive size**. *Science* is evaluating them. Until that evaluation finishes, the exact figures sit under a question mark that only the journal can lift.

So this is two stories at once: an intriguing result about whether facts can move entrenched beliefs, and a live example of the scientific record correcting itself in the open. The point of this piece is to keep them straight.

Belief in a chosen conspiracy, before and after an AI conversation

Study 1: mean belief among the 763 participants who believed their conspiracy, by condition



Reconstructed by The Clean Paper from the public OSF dataset (Costello, Pennycook & Rand, Science 385, eadq1814, 2024).
Dataset under a June 2026 Editorial Expression of Concern; values are provisional, not exact published figures.
Markers = group means; whiskers = 95% CI; shaded band = treatment-control gap.

In the study, a three-round AI conversation that rebutted a participant's own stated evidence was reported to cut belief in that person's chosen conspiracy by about 20% on average; unlike a control chat on an unrelated topic, the drop was still measurable at 10 days and 2 months. The reported effect was durable but partial: most participants still believed the conspiracy afterward — a reduction, not a cure. The paper is currently under an Editorial Expression of Concern (*Science*, June 2026) over reporting inconsistencies and an error in the public dataset; the authors say a corrected analysis preserves the effect's direction, significance and size, while *Science* continues to evaluate. This chart is reconstructed by The Clean Paper from that public dataset, so the values shown are provisional, not the paper's final figures.

What the authors did

- Ran two experiments with 2190 participants who each named — in their own words — a conspiracy theory they believed, and the evidence they thought supported it.
- Had each participant hold a three-round conversation with GPT-4 Turbo. In the **treatment** condition the AI was prompted to rebut the participant's specific evidence; in the **control** condition it discussed an unrelated topic.
- Measured belief in the chosen conspiracy before and immediately after the conversation, and then re-contacted participants at **10 days and 2 months**.
- Had a professional fact-checker evaluate the accuracy of all 128 factual claims the AI made across a sample.
- Checked whether the effect spilled over to other, unrelated conspiracies — and, as a specificity test, whether it also pushed down belief in conspiracies that happen to be **true**.

What they found

- **A roughly 20% average reduction** in belief in the chosen conspiracy in the treatment group relative to control (study 1: 95% confidence interval [13.8, 19.7] points on a 100-point scale, $P < 0.001$).
- **It lasted.** In the treatment group the drop didn't fade: belief stayed near its just-after-the-chat level at 10 days and two months rather than creeping back, while control stayed higher throughout — the paper describes the effect on the chosen conspiracy as persisting undiminished for at least two months, not a momentary wobble.
- **It generalized** across the range of conspiracies people named, and held even for participants whose belief was initially strong.
- **The AI's claims were accurate** in this setting: 127 of 128 (99.2%) rated true, 1 (0.8%) misleading, none false.
- **It was specific**, not blanket doubt: the intervention did **not** reduce belief in true conspiracies, suggesting it moved unsupported beliefs rather than making people cynical about everything.
- **It spilled over modestly** — belief in other, unrelated conspiracies fell by around 12%, and participants reported greater intention to push back on conspiracy claims. The main effect held in study 2, and pointed the same way in a smaller sample that had no control group (weaker evidence, but consistent).

What this does not prove

- It does **not** stand un-questioned right now. The paper is under an Editorial Expression of Concern for data-handling and reproducibility problems, and the corrected numbers are still being evaluated by *Science*. Treat the specific values as provisional until that review lands.
- It does **not** show AI "cures" conspiracy belief. A ~20% average reduction is a meaningful shift, not erasure — most participants still believed the theory afterwards, just less.
- It is **not** a real-world deployment result. Participants opted into a study and engaged with the AI in good faith; in the wild, the people most committed to a conspiracy are the least likely to seek out a chatbot that will argue with them.
- The 99.2% accuracy is **specific to this task, model, and careful prompting** — not a general guarantee that language models state true things. The authors are explicit that the same personalized persuasive power could push *false* beliefs if it were aimed that way.
- "Durable" here means **two months**, which is impressive for a single conversation, but is not the same as permanent.

How strong is the evidence

- **The design is a real strength.** Controlled treatment-versus-control experiments, a clean placebo-style control, replication across two studies and an independent sample, durability follow-ups,

and an explicit accuracy check on the AI's own claims — this is more careful than most persuasion research, and the direction of the effect is consistent everywhere it was tested.

- **But the Expression of Concern is not a footnote.** Being able to regenerate the reported numbers from the released data and code is part of what makes a result trustworthy, and that is precisely what broke — screening-criteria inconsistencies and duplicated rows from a code-merging error. The authors' statement that a corrected pipeline preserves the result is encouraging and plausible, but it is the authors' own account, not yet an independent verdict. *Science*'s evaluation is the thing to wait for.
- **The honest status is "flagged," not "debunked" or "confirmed."** A well-built, striking study whose exact numbers are under formal review. That is an uncomfortable place to leave a good story, and it is the accurate one.

Why it matters

For years, an influential view held that conspiracy believers are driven by psychological needs and identity, and therefore cannot be argued out of their beliefs with facts. A durable, evidence-based reduction — if it holds up — is a meaningful counterweight to that pessimism: it suggests the problem was partly that generic debunking never engaged the *specific* evidence a believer actually cites, something an LLM can do at scale.

It also matters as a case study in how science is supposed to work. The lesson of the Expression of Concern is not that celebrated results are worthless; it is that errors get surfaced, data gets re-examined, and claims get re-checked in public. That machinery running is a feature, not a scandal — and it is the reason the right posture toward this result is patience rather than either applause or dismissal.

And it cuts both ways on AI. The same capacity to deflate a false belief with a tailored argument could inflate one just as effectively. The technique is neutral; only the aim is not.

Clean summary

A 2024 study found that a three-round conversation with GPT-4 Turbo reduced people's belief in a conspiracy theory they held by about 20%, an effect that persisted for two months and generalized across conspiracy types, with the AI's factual claims rated overwhelmingly accurate. In June 2026 the paper was placed under an Editorial Expression of Concern for data-handling and reproducibility problems; the authors report that a corrected analysis preserves the finding in direction, significance, and size, and *Science* is still evaluating. Read it as a genuinely interesting, carefully built result that is currently under review — not settled, not debunked. The most honest sentence about it is the one the process itself is writing: check it again.

Sources

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) — prepended to the article PDF
https://doi.org/10.1126/science.adq1814