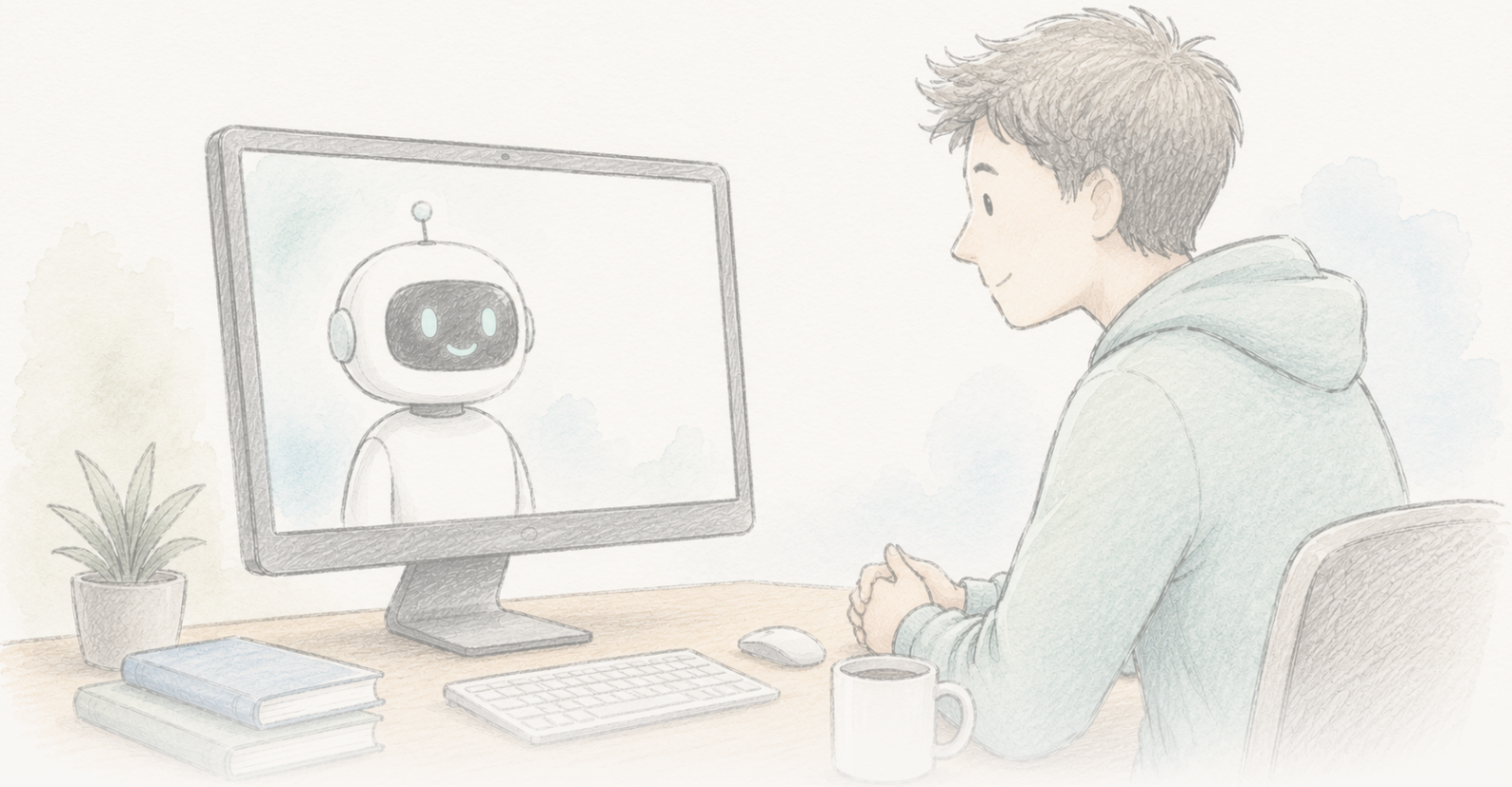


The Clean Paper

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

Um chatbot pareceu reduzir crenças conspirativas durante meses

Um estudo de 2024 concluiu que uma conversa em três rondas com GPT-4 Turbo reduzia em cerca de 20% a crença das pessoas numa teoria da conspiração em que acreditavam, efeito que durou dois meses, se generalizou a diferentes tipos de conspiração e assentou em afirmações da IA avaliadas como esmagadoramente exatas. Em junho de 2026, o artigo recebeu uma Editorial Expression of Concern devido a problemas de tratamento de dados e reprodutibilidade; os autores dizem que uma análise corrigida preserva o resultado na direção, significância e dimensão, e a Science continua a avaliá-lo. Um resultado genuinamente interessante e cuidadosamente construído — atualmente sob revisão, nem encerrado nem 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

A Science associou a este artigo uma Editorial Expression of Concern enquanto avalia questões de tratamento de dados e reprodutibilidade. Não é uma retratação nem uma conclusão de má conduta; significa que o resultado deve ser lido como provisório até a revisão estar resolvida.

Editorial Expression of Concern →

Factos, afinal — e uma sinalização no artigo

Em 2024, um estudo relatou algo que contraria uma ideia confortável e bastante difundida. Dê-se a alguém uma conversa curta, em três rondas, com uma IA — GPT-4 Turbo, instruído para responder à evidência específica que essa pessoa invoca a favor de uma teoria da conspiração em que acredita — e, em média, a crença nessa teoria diminui cerca de 20%. A redução continuava presente dois meses depois. Funcionou com conspirações clássicas (o assassínio de John F. Kennedy, extraterrestres, os Illuminati) e com temas atuais (COVID-19, eleições presidenciais dos EUA de 2020), mesmo entre pessoas cuja crença era forte e ligada à identidade. Quando um verificador profissional de factos avaliou 128 afirmações produzidas pela IA, 127 eram verdadeiras, uma era enganadora e nenhuma era falsa.

A manchete que circulou dizia que *é possível* tirar as pessoas do «buraco do coelho» através da conversa — que quem acredita em conspirações não está fora do alcance da evidência, apenas precisa da evidência certa, apresentada de forma suficientemente específica. É uma afirmação genuinamente interessante, e o estudo foi construído com mais cuidado do que grande parte da investigação sobre persuasão.

Depois, em junho de 2026, a *Science* colocou no artigo uma **Editorial Expression of Concern**. É a parte que a maioria das recontagens irá saltar — e é precisamente a parte que determina quanto peso o resultado consegue suportar neste momento.

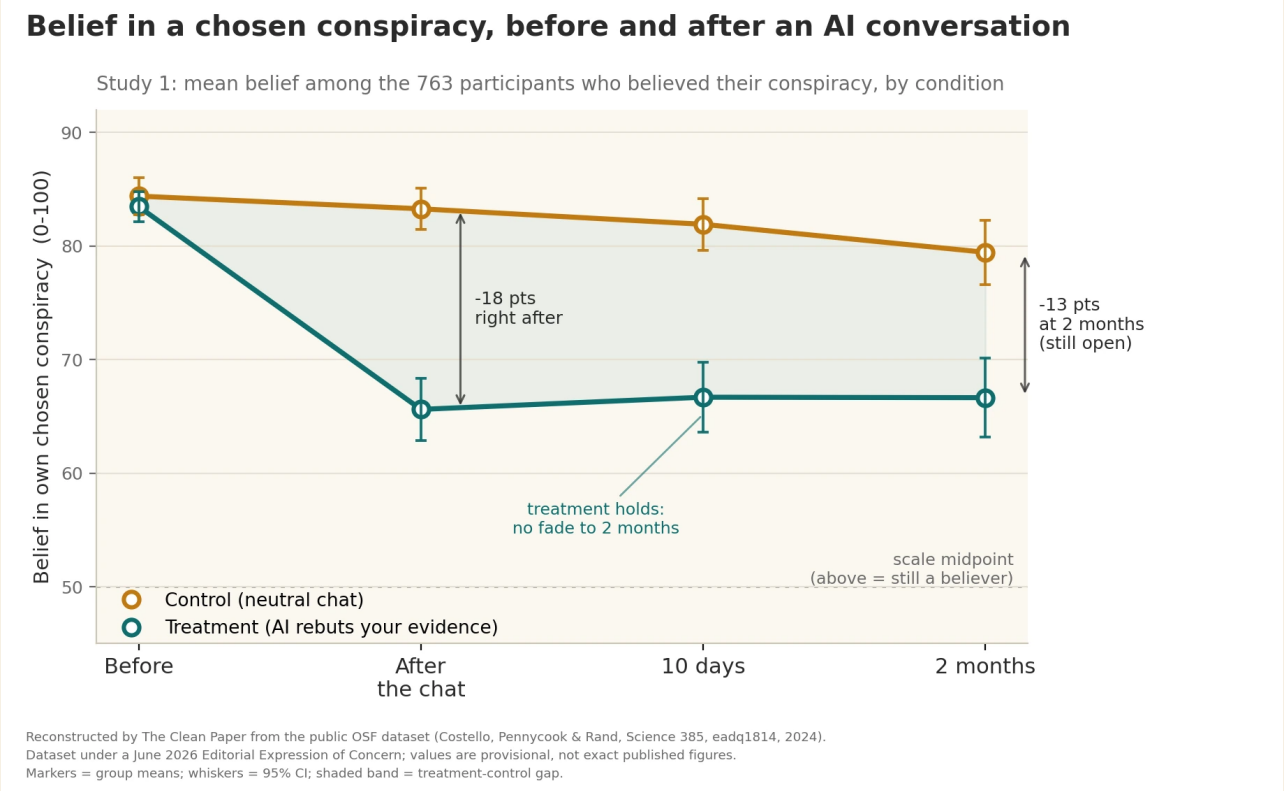
O que é — e não é — uma Expression of Concern

Uma Editorial Expression of Concern é um aviso formal que uma revista associa a um artigo para alertar os leitores de que foram levantadas questões enquanto essas questões ainda estão a ser avaliadas. **Não** é uma retratação (o artigo não foi retirado) e, por si só, não constitui uma conclusão de má conduta. Significa: leia-se o artigo tendo esta ressalva em mente, porque o registo poderá mudar. Neste caso, depois de serem informados dos problemas, os autores investigaram, comunicaram os detalhes à Science e forneceram uma análise corrigida.

O que foi assinalado é específico. Os autores foram informados de inconsistências na aplicação dos critérios de seleção de participantes entre o manuscrito e o código de análise publicado, e de que o conjunto de dados público continha linhas estranhas introduzidas por um erro ao juntar código — problemas que dificultavam reproduzir alguns dos números publicados a partir dos materiais disponibilizados. Os autores entregaram à *Science* um pipeline de análise corrigido e um conjunto de

resultados atualizados, que afirmam coincidir com os originais **na direção, significância estatística e dimensão substantiva**. A *Science* está a avaliá-los. Até essa avaliação terminar, os valores exatos ficam sob um ponto de interrogação que apenas a revista pode retirar.

Portanto, são duas histórias ao mesmo tempo: um resultado intrigante sobre se factos conseguem alterar crenças entrincheiradas e um exemplo vivo do registo científico a corrigir-se em público. O objetivo deste texto é manter as duas coisas separadas.



No estudo, uma conversa em três rondas com uma IA que refutava a evidência indicada pelo próprio participante foi relatada como reduzindo em cerca de 20%, em média, a crença na conspiração escolhida; ao contrário de uma conversa de controlo sobre um tema não relacionado, a redução continuava mensurável aos 10 dias e aos 2 meses. O efeito relatado era duradouro, mas parcial: a maioria dos participantes continuava a acreditar na conspiração depois — uma redução, não uma cura. O artigo está atualmente sob uma Editorial Expression of Concern (*Science*, junho de 2026) devido a inconsistências de reporte e a um erro no conjunto de dados público; os autores afirmam que uma análise corrigida preserva a direção, significância e dimensão do efeito, enquanto a *Science* continua a avaliar. Este gráfico foi reconstruído pelo The Clean Paper a partir desse conjunto de dados público, pelo que os valores apresentados são provisórios, e não os números finais do artigo.

O que fizeram os autores

- Realizaram duas experiências com 2190 participantes, cada um dos quais indicou — pelas suas próprias palavras — uma teoria da conspiração em que acreditava e a evidência que pensava

sustentá-la.
- Cada participante teve uma conversa de três rondas com o GPT-4 Turbo. Na condição de **tratamento**, a IA recebeu instruções para refutar a evidência específica do participante; na condição de **controlo**, discutiu um tema não relacionado.
- Mediram a crença na conspiração escolhida antes e imediatamente depois da conversa e voltaram a contactar os participantes aos **10 dias e 2 meses**.
- Pediram a um verificador profissional de factos que avaliasse a exatidão de todas as 128 afirmações factuais feitas pela IA numa amostra.
- Verificaram se o efeito se estendia a outras conspirações não relacionadas — e, como teste de especificidade, se também reduzia a crença em conspirações que, por acaso, são **verdadeiras**.

O que encontraram

- **Uma redução média de cerca de 20%** na crença na conspiração escolhida no grupo de tratamento relativamente ao controlo (estudo 1: intervalo de confiança de 95% [13,8; 19,7] pontos numa escala de 100 pontos, $P < 0{,}001$).
- **Durou.** No grupo de tratamento, a redução não desapareceu: a crença manteve-se perto do nível observado imediatamente após a conversa aos 10 dias e dois meses, em vez de voltar a subir, enquanto o controlo permaneceu mais alto durante todo o período — o artigo descreve o efeito sobre a conspiração escolhida como persistindo sem diminuição durante pelo menos dois meses, não como uma oscilação momentânea.
- **Generalizou-se** aos diferentes tipos de conspirações mencionados pelas pessoas e manteve-se mesmo entre participantes cuja crença inicial era forte.
- **As afirmações da IA foram exatas** neste contexto: 127 de 128 (99,2%) foram classificadas como verdadeiras, 1 (0,8%) como enganadora e nenhuma como falsa.
- **Foi específico**, e não um aumento geral de desconfiança: a intervenção **não** reduziu a crença em conspirações verdadeiras, sugerindo que deslocou crenças sem sustentação em vez de tornar as pessoas cínicas acerca de tudo.
- **Houve um pequeno efeito de transferência** — a crença noutras conspirações não relacionadas caiu cerca de 12%, e os participantes relataram maior intenção de contestar afirmações conspirativas. O efeito principal manteve-se no estudo 2 e apontou na mesma direção numa amostra menor sem grupo de controlo (evidência mais fraca, mas consistente).

O que isto não demonstra

- **Neste momento, não é um resultado livre de dúvidas.** O artigo está sob uma Editorial Expression of Concern devido a problemas de tratamento de dados e reprodutibilidade, e os números corrigidos continuam a ser avaliados pela *Science*. Os valores específicos devem ser tratados como provisórios até essa revisão terminar.
- **Não** mostra que a IA «cura» a crença em conspirações. Uma redução média de ~20% é uma mudança significativa, não eliminação — a maioria dos participantes continuou a acreditar na

teoria, apenas com menos intensidade.

- **Não** é um resultado de implementação no mundo real. Os participantes escolheram participar num estudo e interagiram de boa-fé com a IA; fora do laboratório, as pessoas mais comprometidas com uma conspiração são provavelmente as menos inclinadas a procurar um chatbot que as contradiga.
- A exatidão de 99,2% é **específica desta tarefa, modelo e prompting cuidadoso** — não é uma garantia geral de que modelos de linguagem dizem coisas verdadeiras. Os autores são explícitos em dizer que o mesmo poder persuasivo personalizado poderia promover crenças *falsas* se fosse orientado nesse sentido.
- «Duradouro» significa aqui **dois meses**, o que é impressionante para uma única conversa, mas não é o mesmo que permanente.

Quão forte é a evidência

- **O desenho é um verdadeiro ponto forte.** Experiências controladas tratamento-versus-controlo, um controlo limpo ao estilo placebo, replicação em dois estudos e numa amostra independente, acompanhamentos de durabilidade e uma verificação explícita da exatidão das afirmações da própria IA — é mais cuidadoso do que grande parte da investigação em persuasão, e a direção do efeito é consistente onde foi testada.
- **Mas a Expression of Concern não é uma nota de rodapé.** Conseguir regenerar os números publicados a partir dos dados e código disponibilizados faz parte da confiança num resultado, e foi precisamente isso que falhou — inconsistências nos critérios de seleção e linhas duplicadas introduzidas por um erro de fusão de código. A afirmação dos autores de que um pipeline corrigido preserva o resultado é encorajadora e plausível, mas continua a ser o relato dos próprios autores, ainda não um veredicto independente. É a avaliação da *Science* que é preciso esperar.
- **O estado honesto é «sinalizado», não «refutado» nem «confirmado».** Um estudo bem construído e marcante cujos números exatos estão sob revisão formal. É um lugar desconfortável para deixar uma boa história — e é o lugar correto.

Porque importa

Durante anos, uma perspetiva influente sustentou que quem acredita em conspirações é movido por necessidades psicológicas e identidade e, por isso, não pode ser afastado dessas crenças através de factos. Uma redução duradoura baseada em evidência — se resistir — seria um contraponto importante a esse pessimismo: sugeriria que o problema estava em parte no facto de a refutação genérica nunca se envolver com a *evidência específica* que uma pessoa realmente cita, algo que um LLM consegue fazer à escala.

Também importa como estudo de caso sobre a forma como a ciência deve funcionar. A lição da Expression of Concern não é que resultados celebrados não tenham valor; é que os erros emergem, os dados são reexaminados e as afirmações são verificadas novamente em público. Esse mecanismo

a funcionar é uma característica, não um escândalo — e é por isso que a postura correta perante este resultado é paciência, não aplauso nem rejeição.

E o tema da IA corta nos dois sentidos. A mesma capacidade de enfraquecer uma crença falsa com um argumento personalizado poderia reforçar uma crença falsa com igual eficácia. A técnica é neutra; o objetivo não é.

Resumo claro

Um estudo de 2024 concluiu que uma conversa em três rondas com GPT-4 Turbo reduzia em cerca de 20% a crença das pessoas numa teoria da conspiração em que acreditavam, um efeito que persistia durante dois meses e se generalizava a diferentes tipos de conspiração, com as afirmações factuais da IA avaliadas como esmagadoramente exatas. Em junho de 2026, o artigo recebeu uma Editorial Expression of Concern devido a problemas no tratamento dos dados e na reprodutibilidade; os autores afirmam que uma análise corrigida preserva a descoberta na direção, significância e dimensão, e a *Science* continua a avaliá-la. Deve ser lido como um resultado genuinamente interessante e cuidadosamente construído que está atualmente sob revisão — não encerrado, não refutado. A frase mais honesta sobre ele é aquela que o próprio processo científico está agora a escrever: verifique-se outra vez.

Fontes

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`