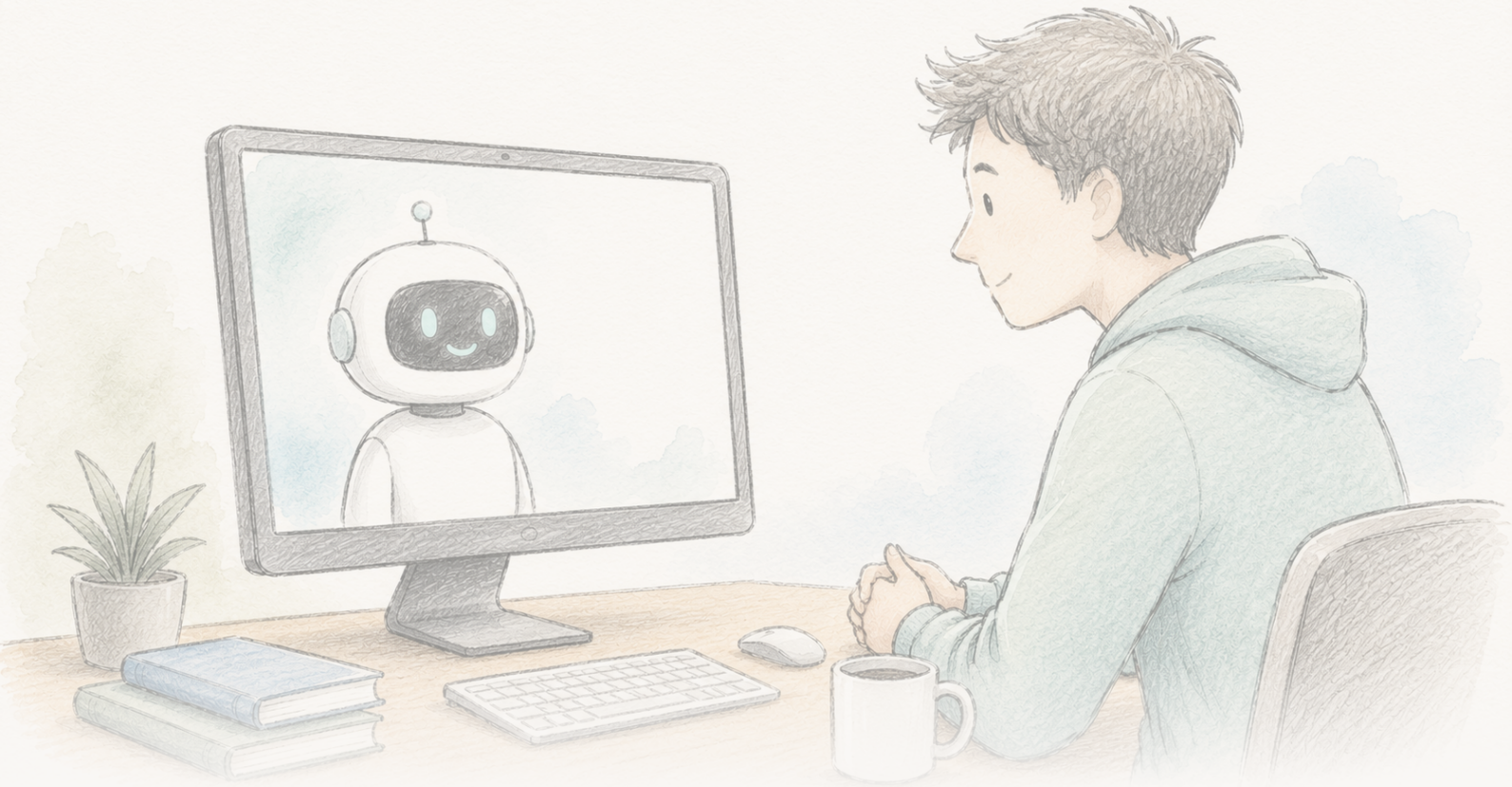


The Clean Paper

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

Um chatbot pareceu reduzir crenças conspiratórias por meses

Um estudo de 2024 encontrou que uma conversa de três rodadas com GPT-4 Turbo reduziu em cerca de 20% a crença das pessoas em uma teoria conspiratória que elas defendiam, um efeito que durou dois meses, se generalizou entre tipos de conspiração e se apoiou em afirmações da IA avaliadas como esmagadoramente corretas. Em junho de 2026, o artigo recebeu uma Editorial Expression of Concern por problemas de manuseio de dados e reprodutibilidade; os autores dizem que uma análise corrigida preserva o achado em direção, significância e tamanho, e a Science ainda está avaliando. Um resultado genuinamente interessante e cuidadosamente construído — atualmente em revisão, nem estabelecido nem derrubado.

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

Editorial Expression of Concern

A Science anexou uma Editorial Expression of Concern a este artigo enquanto avalia questões de manuseio de dados e reprodutibilidade. Isso não é uma retratação nem uma constatação de má conduta; significa que o resultado relatado deve ser lido como provisório até a revisão ser resolvida.

Editorial Expression of Concern →

Fatos, afinal — e uma bandeira no artigo

Em 2024, um estudo relatou algo que contraria uma peça confortável da sabedoria convencional. Dê a alguém uma conversa curta, de três rodadas, com uma IA — GPT-4 Turbo, instruído a responder às evidências específicas que essa pessoa cita para uma teoria da conspiração em que acredita pessoalmente — e, em média, a crença dela nessa teoria cai cerca de 20%. A queda ainda estava presente dois meses depois. Funcionou com conspirações clássicas (o assassinato de John F. Kennedy, alienígenas, Illuminati) e com as atuais (COVID-19, a eleição americana de 2020), mesmo entre pessoas cuja crença era forte e ligada à identidade. Quando um fact-checker profissional avaliou 128 afirmações feitas pela IA, 127 eram verdadeiras, uma era enganosa e nenhuma era falsa.

A manchete que viajou foi que *é possível* conversar alguém para fora da toca do coelho — que crentes em conspirações não estão além do alcance da evidência; eles só precisam da evidência certa, entregue de modo específico o bastante. Essa é uma afirmação genuinamente interessante, e o estudo foi construído com mais cuidado do que a maioria dos trabalhos sobre persuasão.

Então, em junho de 2026, a *Science* publicou uma **Editorial Expression of Concern** sobre o artigo. Essa é a parte que a maioria dos relatos vai pular, e é exatamente a parte que decide quanto peso o resultado pode carregar agora.

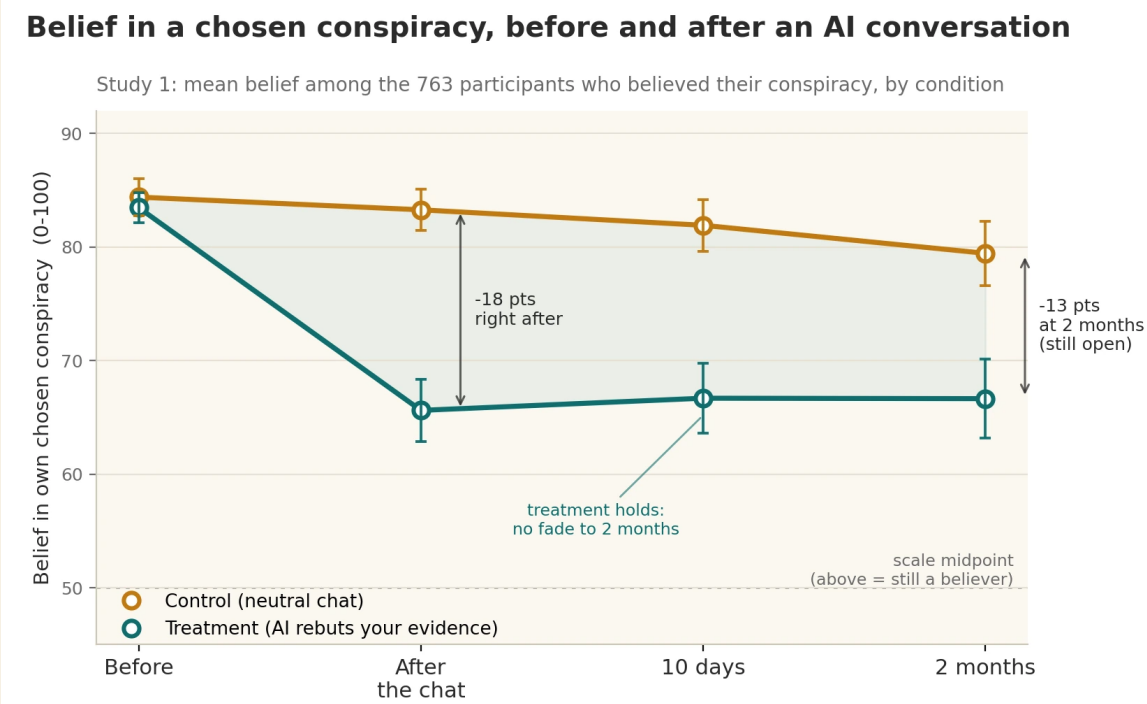
O que é — e o que não é — uma Expression of Concern

Uma Editorial Expression of Concern é uma bandeira formal que uma revista anexa a um artigo para alertar leitores de que questões foram levantadas enquanto essas questões ainda estão sendo apuradas. **Não** é uma retratação (o artigo não foi retirado) e não é, por si só, uma constatação de má conduta. Significa: leia com esse caveat em mente, porque o registro pode mudar. Aqui, depois de serem informados dos problemas, os autores investigaram, relataram os detalhes à Science e forneceram uma análise corrigida.

O que foi sinalizado é específico. Os autores foram alertados sobre inconsistências na aplicação dos critérios de triagem de participantes entre o manuscrito e o código de análise publicado, e sobre linhas extras inseridas no conjunto de dados público por um erro de merge de código — problemas que tornavam alguns números relatados difíceis de reproduzir a partir dos materiais liberados. Os autores entregaram à *Science* uma pipeline de análise corrigida e um conjunto de resultados atualizados, que eles dizem combinar com o original **em direção, significância estatística e tamanho**

substantivo. A *Science* está avaliando. Até essa avaliação terminar, os números exatos ficam sob um ponto de interrogação que só a revista pode levantar.

Então são duas histórias ao mesmo tempo: um resultado intrigante sobre se fatos podem mover crenças arraigadas, e um exemplo vivo do registro científico se corrigindo em público. O ponto deste texto é mantê-las separadas.



No estudo, uma conversa de três rodadas com IA que rebatia as evidências declaradas pelo próprio participante foi relatada como capaz de reduzir a crença na conspiração escolhida em cerca de 20% em média; ao contrário de uma conversa de controle sobre tema não relacionado, a queda ainda era mensurável em 10 dias e 2 meses. O efeito relatado foi durável, mas parcial: a maioria dos participantes ainda acreditava na conspiração depois — uma redução, não uma cura. O artigo está atualmente sob uma Editorial Expression of Concern (*Science*, junho de 2026) por inconsistências de relato e um erro no conjunto de dados público; os autores dizem que uma análise corrigida preserva direção, significância e tamanho do efeito, enquanto a *Science* continua avaliando. Este gráfico foi reconstruído pelo The Clean Paper a partir desse conjunto de dados público, portanto os valores mostrados são provisórios, não os números finais do artigo.

O que os autores fizeram

- Rodaram dois experimentos com 2190 participantes que nomearam — com suas próprias palavras — uma teoria da conspiração em que acreditavam e a evidência que achavam sustentá-la.
- Fizeram cada participante manter uma conversa de três rodadas com o GPT-4 Turbo. Na condição

de **tratamento**, a IA foi instruída a rebater a evidência específica do participante; na condição de **controle**, discutiu um tema não relacionado.
- Mediram a crença na conspiração escolhida antes e imediatamente depois da conversa, e então recontataram participantes em **10 dias e 2 meses**.
- Pediram a um fact-checker profissional que avaliasse a precisão de todas as 128 afirmações factuais feitas pela IA numa amostra.
- Verificaram se o efeito transbordava para outras conspirações não relacionadas — e, como teste de especificidade, se também reduzia crença em conspirações que por acaso eram **verdadeiras**.

O que eles encontraram

- **Uma redução média de cerca de 20%** na crença na conspiração escolhida no grupo de tratamento em relação ao controle (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 queda não desapareceu: a crença ficou perto do nível logo após a conversa em 10 dias e em dois meses, em vez de voltar a subir, enquanto o controle permaneceu mais alto — o artigo descreve o efeito sobre a conspiração escolhida como persistente, sem diminuição, por pelo menos dois meses.
- **Generalizou** pelo conjunto de conspirações nomeadas pelas pessoas e se manteve até para participantes cuja crença inicial era forte.
- **As afirmações da IA foram precisas** nesse cenário: 127 de 128 (99,2%) avaliadas como verdadeiras, 1 (0,8%) enganosa, nenhuma falsa.
- **Foi específico**, não uma dúvida generalizada: a intervenção **não** reduziu crença em conspirações verdadeiras, sugerindo que moveu crenças sem suporte em vez de tornar as pessoas cínicas sobre tudo.
- **Transbordou modestamente** — a crença em outras conspirações não relacionadas caiu cerca de 12%, e participantes relataram maior intenção de contestar alegações conspiratórias. O efeito principal se manteve no estudo 2 e apontou na mesma direção numa amostra menor sem grupo de controle (evidência mais fraca, mas consistente).

O que isso não prova

- **Não** fica sem questionamento agora. O artigo está sob uma Editorial Expression of Concern por problemas de manuseio de dados e reprodutibilidade, e os números corrigidos ainda estão sendo avaliados pela *Science*. Trate os valores específicos como provisórios até essa revisão chegar.
- **Não** mostra que IA "cura" crença conspiratória. Uma redução média de ~20% é uma mudança significativa, não apagamento — a maioria dos participantes ainda acreditava na teoria depois, só menos.
- **Não** é um resultado de implantação no mundo real. Participantes aceitaram entrar num estudo e se envolveram de boa-fé com a IA; no mundo aberto, as pessoas mais comprometidas com uma conspiração são as menos propensas a procurar um chatbot que vá argumentar contra elas.

- A precisão de 99,2% é **específica a esta tarefa, modelo e prompting cuidadoso** — não uma garantia geral de que modelos de linguagem dizem coisas verdadeiras. Os autores são explícitos: o mesmo poder persuasivo personalizado poderia empurrar crenças *falsas* se fosse dirigido para isso.
- "Durável" aqui significa **dois meses**, o que é impressionante para uma conversa única, mas não é o mesmo que permanente.

Quão forte é a evidência

- **O desenho é uma força real.** Experimentos controlados tratamento-versus-controle, um controle limpo em estilo placebo, replicação em dois estudos e uma amostra independente, acompanhamentos de durabilidade e uma checagem explícita da precisão das afirmações da IA — isso é mais cuidadoso que a maioria das pesquisas de persuasão, e a direção do efeito é consistente onde foi testada.
- **Mas a Expression of Concern não é nota de rodapé.** Conseguir regenerar os números relatados a partir dos dados e do código liberados faz parte do que torna um resultado confiável, e foi exatamente isso que quebrou — inconsistências nos critérios de triagem e linhas duplicadas por erro de merge de código. A afirmação dos autores de que uma pipeline corrigida preserva o resultado é encorajadora e plausível, mas é a própria versão dos autores, ainda não um veredito independente. A avaliação da *Science* é o que deve ser aguardado.
- **O status honesto é "sinalizado", não "derrubado" nem "confirmado".** Um estudo bem construído e chamativo cujos números exatos estão sob revisão formal. É um lugar desconfortável para deixar uma boa história, e é o lugar correto.

Por que importa

Por anos, uma visão influente sustentou que pessoas que acreditam em conspirações são movidas por necessidades psicológicas e identidade, e portanto não podem ser convencidas por fatos. Uma redução durável baseada em evidências — se sobreviver — é um contrapeso relevante a esse pessimismo: sugere que o problema era em parte que o debunking genérico nunca lidava com a *evidência específica* que a pessoa realmente cita, algo que um LLM consegue fazer em escala.

Também importa como estudo de caso de como a ciência deve funcionar. A lição da Expression of Concern não é que resultados celebrados sejam inúteis; é que erros aparecem, dados são reexaminados e alegações são rechecadas em público. Essa maquinaria funcionando é uma qualidade, não um escândalo — e é por isso que a postura correta diante deste resultado é paciência, não aplauso nem descarte.

E isso corta dos dois lados sobre IA. A mesma capacidade de reduzir uma crença falsa com um argumento sob medida poderia inflar uma com a mesma eficácia. A técnica é neutra; o objetivo, não.

Resumo limpo

Um estudo de 2024 encontrou que uma conversa de três rodadas com GPT-4 Turbo reduziu a crença das pessoas em uma teoria da conspiração que elas defendiam em cerca de 20%, um efeito que persistiu por dois meses e se generalizou entre tipos de conspiração, com as afirmações factuais da IA avaliadas como esmagadoramente corretas. Em junho de 2026, o artigo recebeu uma Editorial Expression of Concern por problemas de manuseio de dados e reprodutibilidade; os autores relatam que uma análise corrigida preserva o achado em direção, significância e tamanho, e a *Science* ainda está avaliando. Leia como um resultado genuinamente interessante e cuidadosamente construído que está atualmente sob revisão — não estabelecido, não derrubado. A frase mais honesta sobre ele é a que o próprio processo está escrevendo: verifique de novo.

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