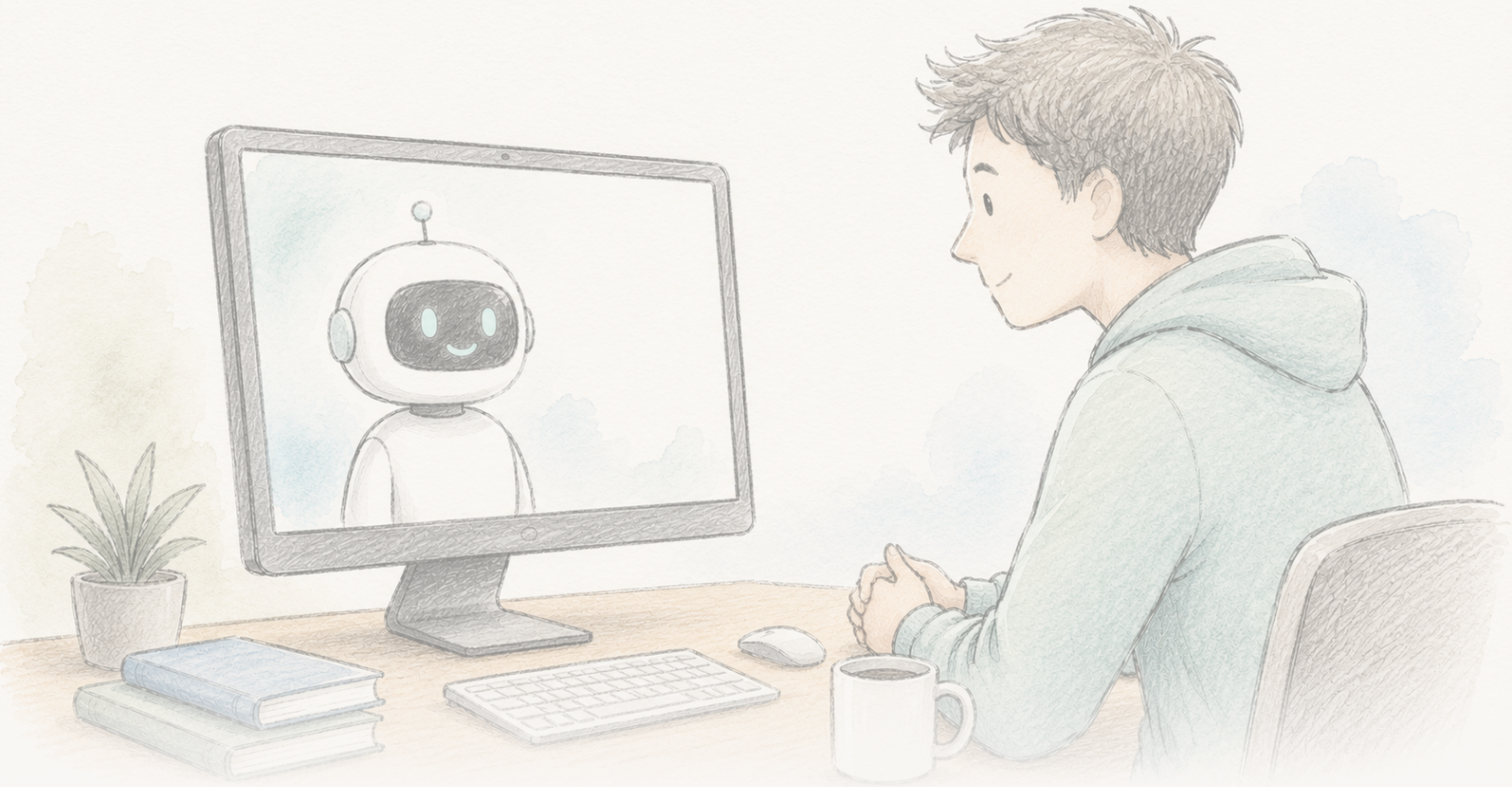


The Clean Paper

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

Một chatbot dường như làm giảm niềm tin vào thuyết âm mưu trong nhiều tháng

Một nghiên cứu năm 2024 cho thấy cuộc trò chuyện ba lượt với GPT-4 Turbo làm giảm khoảng 20% niềm tin của người tham gia vào một thuyết âm mưu họ đang tin; hiệu ứng kéo dài hai tháng, xuất hiện trên nhiều loại thuyết âm mưu và dựa trên các tuyên bố của AI được đánh giá gần như hoàn toàn chính xác. Tháng 6/2026, bài báo bị đặt dưới Editorial Expression of Concern vì các vấn đề xử lý dữ liệu và khả năng tái tạo; các tác giả nói phân tích đã sửa vẫn giữ hướng, ý nghĩa thống kê và độ lớn của kết quả, còn Science vẫn đang đánh giá. Một kết quả thật sự thú vị và được xây dựng cẩn thận — hiện vẫn đang được xem xét, chưa ngã ngũ cũng chưa bị bác bỏ.

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 đã gắn Editorial Expression of Concern lên bài báo này trong khi đánh giá các câu hỏi về xử lý dữ liệu và khả năng tái tạo. Đây không phải rút bài và cũng không phải kết luận có hành vi sai trái; nó có nghĩa kết quả được báo cáo nên được xem là tạm thời cho đến khi quá trình đánh giá được giải quyết.

Editorial Expression of Concern →

Sau cùng vẫn là sự thật — và một lá cờ cảnh báo trên bài báo

Năm 2024, một nghiên cứu báo cáo điều đi ngược một quan niệm khá phổ biến. Cho một người trò chuyện ngắn ba lượt với AI — GPT-4 Turbo, được nhắc để trả lời đúng những bằng chứng cụ thể mà người đó viện dẫn cho một thuyết âm mưu họ thật sự tin — thì trung bình niềm tin của họ vào thuyết ấy giảm khoảng 20%. Hai tháng sau mức giảm vẫn còn. Hiệu ứng xuất hiện ở cả các thuyết âm mưu kinh điển (vụ ám sát John F. Kennedy, người ngoài hành tinh, Illuminati) lẫn các thuyết đương thời (COVID-19, bầu cử Mỹ 2020), kể cả ở những người có niềm tin mạnh và gắn với bản sắc. Khi một chuyên gia kiểm chứng thông tin chuyên nghiệp chấm 128 tuyên bố do AI đưa ra, 127 được đánh giá đúng, một gây hiểu lầm và không có tuyên bố nào sai.

Tiêu đề lan truyền là người ta *có thể* được nói chuyện để thoát khỏi "hang thỏ" — rằng người tin thuyết âm mưu không nằm ngoài tầm với của bằng chứng; họ chỉ cần đúng bằng chứng, được đưa ra đủ cụ thể. Đó là một tuyên bố thực sự thú vị, và nghiên cứu được thiết kế cẩn thận hơn phần lớn công trình về thuyết phục.

Rồi tháng 6/2026, *Science* gắn một **Editorial Expression of Concern** lên bài báo. Đây là phần mà nhiều lần kể lại sẽ bỏ qua, nhưng cũng chính là phần quyết định kết quả hiện có thể gánh bao nhiêu trọng lượng.

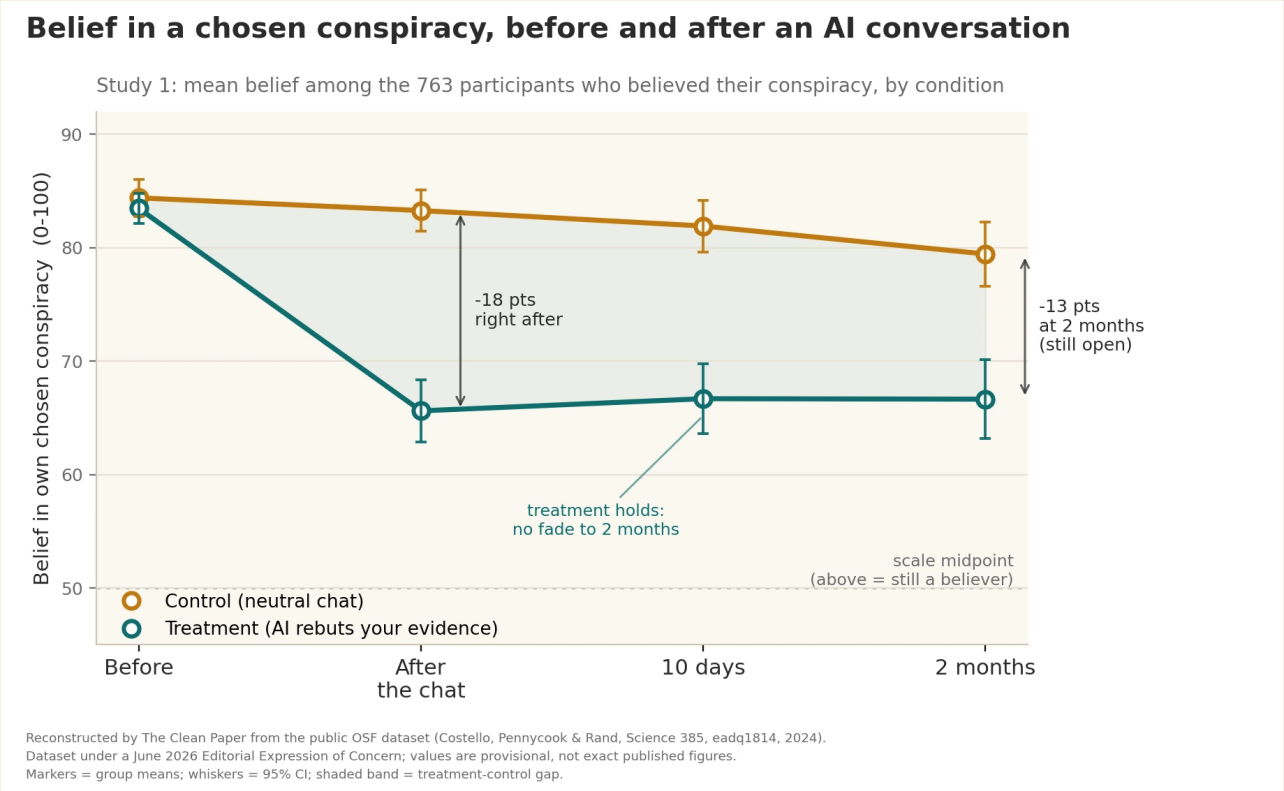
Editorial Expression of Concern là gì — và không phải là gì

Editorial Expression of Concern là một cảnh báo chính thức mà tạp chí gắn vào một bài báo để báo cho độc giả rằng đã có những câu hỏi được nêu ra trong khi các câu hỏi ấy vẫn đang được xem xét. Nó **không** phải một bài bị rút (*retraction*), và bản thân nó cũng không phải kết luận có hành vi sai trái. Ý nghĩa là: hãy đọc kết quả với lưu ý này trong đầu, vì hồ sơ khoa học có thể còn thay đổi. Trong trường hợp này, sau khi được thông báo về các vấn đề, các tác giả đã điều tra, báo chi tiết cho *Science* và cung cấp một phân tích đã sửa.

Vấn đề được nêu ra khá cụ thể. Các tác giả được thông báo rằng tiêu chí sàng lọc người tham gia đã được áp dụng không nhất quán giữa bản thảo và mã phân tích công bố, và bộ dữ liệu công khai có các hàng thừa bị chèn vào do lỗi gộp mã — những lỗi khiến một số con số được

báo cáo khó tái tạo từ tài liệu phát hành. Các tác giả đã đưa cho *Science* một pipeline phân tích đã sửa và một tập kết quả cập nhật mà họ nói vẫn khớp với kết quả ban đầu **về hướng, ý nghĩa thống kê và độ lớn thực chất**. *Science* đang đánh giá chúng. Cho đến khi đánh giá hoàn tất, các con số chính xác vẫn mang một dấu hỏi mà chỉ tạp chí mới có thể gỡ bỏ.

Vì vậy đây đồng thời là hai câu chuyện: một kết quả hấp dẫn về việc sự thật có thể làm lung lay niềm tin ăn sâu đến đâu, và một ví dụ đang diễn ra về cách hồ sơ khoa học tự sửa mình công khai. Mục tiêu của bài này là giữ hai chuyện đó tách bạch.



Trong nghiên cứu, một cuộc trò chuyện AI ba lượt phản bác chính các bằng chứng do người tham gia nêu ra được báo cáo là làm giảm trung bình khoảng 20% niềm tin vào thuyết âm mưu mà người đó chọn; khác với một cuộc trò chuyện đối chứng về chủ đề không liên quan, mức giảm vẫn đo được sau 10 ngày và 2 tháng. Hiệu ứng được báo cáo là bền nhưng chỉ một phần: phần lớn người tham gia vẫn tin thuyết âm mưu sau đó — mức độ tin giảm, không phải được "chữa khỏi". Bài báo hiện đang chịu một Editorial Expression of Concern (*Science*, tháng 6/2026) vì các bất nhất trong báo cáo và một lỗi trong bộ dữ liệu công khai; các tác giả nói phân tích đã sửa vẫn giữ hướng, ý nghĩa và độ lớn của hiệu ứng, trong khi *Science* tiếp tục đánh giá. Biểu đồ này được The Clean Paper dựng lại từ bộ dữ liệu công khai đó, nên các giá trị hiển thị là tạm thời, không phải con số cuối cùng của bài báo.

Các tác giả đã làm gì

- Thực hiện hai thí nghiệm với 2.190 người tham gia; mỗi người tự nêu, bằng lời của mình, một thuyết âm mưu họ tin và bằng chứng họ cho là ủng hộ nó.
- Cho mỗi người trò chuyện ba lượt với GPT-4 Turbo. Trong điều kiện **can thiệp**, AI được yêu cầu phản bác bằng chứng cụ thể của người tham gia; trong điều kiện **đối chứng**, AI thảo luận một chủ đề không liên quan.
- Đo niềm tin vào thuyết âm mưu đã chọn trước và ngay sau cuộc trò chuyện, rồi liên hệ lại sau **10 ngày và 2 tháng**.
- Nhờ một chuyên gia kiểm chứng thông tin chuyên nghiệp đánh giá độ chính xác của toàn bộ 128 tuyên bố thực tế do AI đưa ra trong một mẫu.
- Kiểm tra xem hiệu ứng có lan sang các thuyết âm mưu khác, không liên quan hay không — và như một phép thử tính đặc hiệu, liệu nó có làm giảm cả niềm tin vào những "âm mưu" hóa ra **có thật** hay không.

Họ tìm thấy gì

- **Mức giảm trung bình khoảng 20%** trong niềm tin vào thuyết âm mưu đã chọn ở nhóm can thiệp so với nhóm đối chứng (nghiên cứu 1: khoảng tin cậy 95% [13,8; 19,7] điểm trên thang 100 điểm, $P < 0,001$).
- **Hiệu ứng kéo dài.** Ở nhóm can thiệp, mức giảm không biến mất: niềm tin vẫn gần mức ngay sau cuộc trò chuyện ở mốc 10 ngày và hai tháng, thay vì từ từ trở lại, trong khi nhóm đối chứng vẫn cao hơn suốt thời gian theo dõi — bài báo mô tả hiệu ứng lên thuyết âm mưu được chọn là kéo dài không suy giảm ít nhất hai tháng, chứ không phải một dao động thoáng qua.
- **Nó xuất hiện rộng** trên nhiều loại thuyết âm mưu mà người tham gia nêu ra, kể cả ở những người ban đầu tin rất mạnh.
- **Các tuyên bố của AI chính xác trong bối cảnh này:** 127/128 (99,2%) được đánh giá đúng, 1 (0,8%) gây hiểu lầm, không có tuyên bố nào sai.
- **Hiệu ứng có tính đặc hiệu**, không phải hoài nghi tràn lan: can thiệp **không** làm giảm niềm tin vào các âm mưu có thật, gợi ý rằng nó tác động lên những niềm tin thiếu cơ sở thay vì khiến mọi người hoài nghi tất cả.
- **Có một mức lan tỏa khiêm tốn:** niềm tin vào các thuyết âm mưu khác, không liên quan giảm khoảng 12%, và người tham gia báo ý định phản bác các tuyên bố âm mưu nhiều hơn. Hiệu ứng chính được lặp lại trong nghiên cứu 2 và chỉ cùng hướng trong một mẫu nhỏ hơn không có nhóm đối chứng (bằng chứng yếu hơn nhưng nhất quán).

Điều nghiên cứu này không chứng minh

- Hiện nó **không đứng ở trạng thái không bị nghi vấn**. Bài báo đang chịu Editorial Expression of Concern vì các vấn đề xử lý dữ liệu và khả năng tái tạo, và các con số đã sửa vẫn đang được *Science* đánh giá. Hãy xem những giá trị cụ thể là tạm thời cho đến khi quá trình ấy kết thúc.
- Nó **không** cho thấy AI "chữa khỏi" niềm tin âm mưu. Giảm trung bình khoảng 20% là thay đổi đáng kể, không phải xóa bỏ — phần lớn người tham gia vẫn tin thuyết ấy sau đó, chỉ là tin ít hơn.
- Đây **không** phải kết quả triển khai ngoài đời thực. Người tham gia tự nguyện vào nghiên cứu và tương tác với AI một cách thiện chí; ngoài đời, những người gắn bó nhất với một thuyết âm mưu có thể chính là những người ít muốn tìm một chatbot để tranh luận với mình nhất.
- Độ chính xác 99,2% **chỉ thuộc nhiệm vụ, mô hình và cách prompting cẩn thận này** — không phải bảo đảm chung rằng mô hình ngôn ngữ nói đúng sự thật. Các tác giả nói rõ cùng năng lực thuyết phục cá nhân hóa đó cũng có thể đẩy người ta về phía niềm tin *sai* nếu được dùng theo hướng ấy.
- "Bền" ở đây nghĩa là **hai tháng**, rất ấn tượng cho một cuộc trò chuyện duy nhất nhưng không đồng nghĩa với vĩnh viễn.

Bằng chứng mạnh đến đâu

- **Thiết kế là một điểm mạnh thật sự.** Thí nghiệm có can thiệp và đối chứng, một đối chứng kiểu placebo khá sạch, lặp lại qua hai nghiên cứu và một mẫu độc lập, theo dõi độ bền, cùng kiểm tra riêng độ chính xác của chính các tuyên bố AI — cẩn thận hơn phần lớn nghiên cứu thuyết phục, và hướng của hiệu ứng nhất quán ở mọi nơi nó được kiểm tra.
- **Nhưng Expression of Concern không phải một chú thích nhỏ.** Khả năng tái tạo các con số được báo cáo từ dữ liệu và mã công khai là một phần của độ tin cậy, và đây chính là chỗ đã hỏng — bất nhất trong tiêu chí sàng lọc và các hàng dữ liệu trùng/thừa từ lỗi gộp mã. Tuyên bố của các tác giả rằng pipeline đã sửa vẫn giữ kết quả là đáng khích lệ và có thể hợp lý, nhưng đó vẫn là lời của chính các tác giả, chưa phải phán quyết độc lập. Điều cần chờ là đánh giá của *Science*.
- **Trạng thái trung thực là "đang bị gắn cờ", không phải "đã bác bỏ" hay "đã xác nhận".** Một nghiên cứu được xây dựng công phu và gây chú ý, nhưng các con số chính xác đang được xem xét chính thức. Đó là một nơi khó chịu để bỏ lại một câu chuyện hay, nhưng là nơi chính xác.

Vì sao nó quan trọng

Trong nhiều năm, một quan điểm có ảnh hưởng cho rằng người tin thuyết âm mưu bị thúc đẩy bởi nhu cầu tâm lý và bản sắc, vì vậy không thể được thuyết phục bằng sự thật. Một mức giảm bền dựa trên bằng chứng — nếu nó đứng vững — là đối trọng đáng kể với sự bi quan ấy: nó gợi ý rằng một phần vấn đề là phản bác chung chung chưa bao giờ chạm vào *đúng bằng chứng cụ thể* mà người tin thật sự viện dẫn, điều mà một LLM có thể làm ở quy mô lớn.

Nó cũng quan trọng như một nghiên cứu tình huống về cách khoa học đáng lẽ phải vận hành. Bài học từ Expression of Concern không phải là các kết quả nổi tiếng đều vô giá trị; mà là lỗi được đưa ra ánh sáng, dữ liệu được xem lại và tuyên bố được kiểm tra lại công khai. Bộ máy ấy đang chạy là một đặc tính tốt, không phải bê bối — và vì vậy thái độ đúng với kết quả này là kiên nhẫn, thay vì vỗ tay hay bác bỏ ngay.

Và với AI, nó có hai mặt. Cùng khả năng làm suy yếu một niềm tin sai bằng lập luận được cá nhân hóa cũng có thể làm một niềm tin sai phồng lên hiệu quả tương tự. Kỹ thuật là trung tính; mục tiêu thì không.

Tóm tắt gọn

Một nghiên cứu năm 2024 cho thấy một cuộc trò chuyện ba lượt với GPT-4 Turbo làm giảm khoảng 20% niềm tin của người tham gia vào thuyết âm mưu họ đang tin, hiệu ứng kéo dài hai tháng và xuất hiện trên nhiều loại thuyết âm mưu, trong khi các tuyên bố thực tế của AI được đánh giá là gần như hoàn toàn chính xác. Tháng 6/2026, bài báo bị đặt dưới Editorial Expression of Concern vì các vấn đề xử lý dữ liệu và khả năng tái tạo; các tác giả báo rằng phân tích sửa đổi vẫn giữ hướng, ý nghĩa thống kê và độ lớn của phát hiện, còn *Science* vẫn đang đánh giá. Hãy đọc đây là một kết quả thật sự thú vị và được thiết kế cẩn thận nhưng hiện còn đang được xem xét — chưa ngã ngũ, cũng chưa bị bác bỏ. Câu trung thực nhất về nó chính là câu mà quy trình khoa học đang viết: hãy kiểm tra lại.

Nguồn

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`