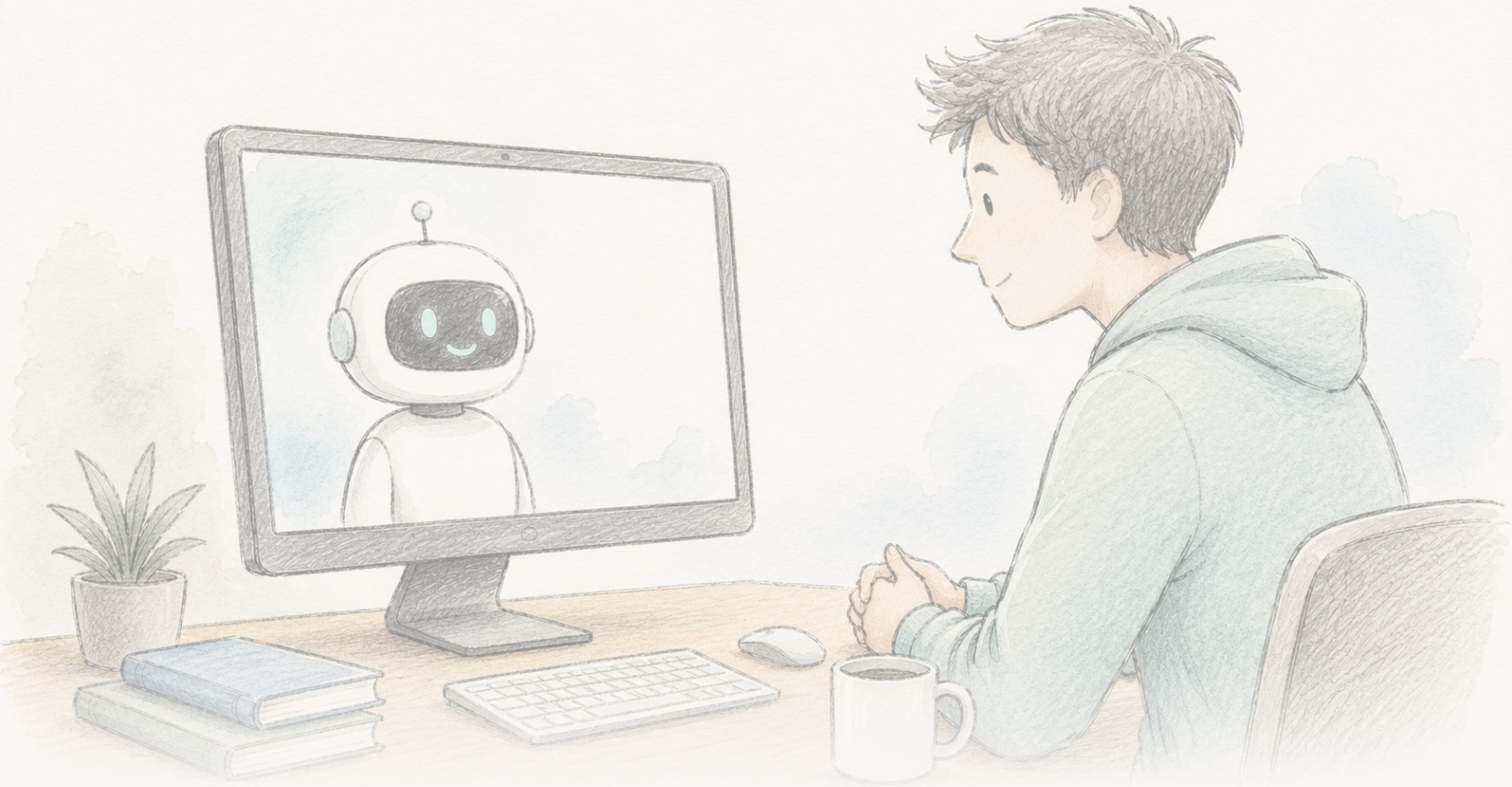


The Clean Paper

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

一個聊天機器人似乎持續減少了陰謀論信念達數月

一項 *2024* 年的研究發現，與 *GPT-4 Turbo* 的三輪對話使人對其所持的一項陰謀論的信念削減了約 *20%*，該效應持續了兩個月，並在各陰謀論類型間泛化，且基於被評定為壓倒性準確的 *AI* 聲明。*2026* 年 *6* 月，該論文因資料處理和可復現性問題被置於一項編輯部關切聲明之下；作者表示一項經修正的分析在方向、顯著性和效應量上保留了該發現，《科學》仍在評估中。一項真正有趣的、構建仔細的——目前正在審查中的結果，既非已解決，也非已被駁斥。

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

《科學》在評估資料處理和可復現性問題期間，已對該論文附上一則編輯部關切聲明。這不是撤稿，也不是不當行為的發現；它意味著報告的結果在其審查解決之前應被視為臨時的。

Editorial Expression of Concern →

事實，終究——以及論文上的一面旗幟

2024 年，一項研究報告了一件與一條舒適的常識常理相悖的發現。讓人與 AI（GPT-4 Turbo，被提示去反駁他們個人相信的一項陰謀論所引用的具體證據）進行一次簡短的三輪對話，平均而言，他們對該理論的信念下降了約 20%。這一下降在兩個月後仍然存在。它跨越了經典陰謀論（甘迺迪遇刺、外星人、光明會）和時政陰謀論（COVID-19、2020 年美國大選），甚至在那些信念強烈且與身份綁定的參與者中也有效。當一名專業事實核查員對 AI 所做的 128 條聲明的樣本進行評估時，127 條為真，1 條為誤導，0 條為假。

廣為流傳的標題是，你可以把人從兔子洞裡拉出來——陰謀論者並非超出證據的觸及範圍，他們只是需要正確的證據，以足夠具體的方式傳遞。這是一個真正有趣的主張，該研究的構建比大多數說服研究更為仔細。

然後，在 2026 年 6 月，《科學》對該論文發布了一則**編輯部關切聲明**。這是大多數複述會跳過的部分，而這恰恰是決定這一結果目前能承受多大分量的部分。

編輯部關切聲明是什麼——以及不是什麼

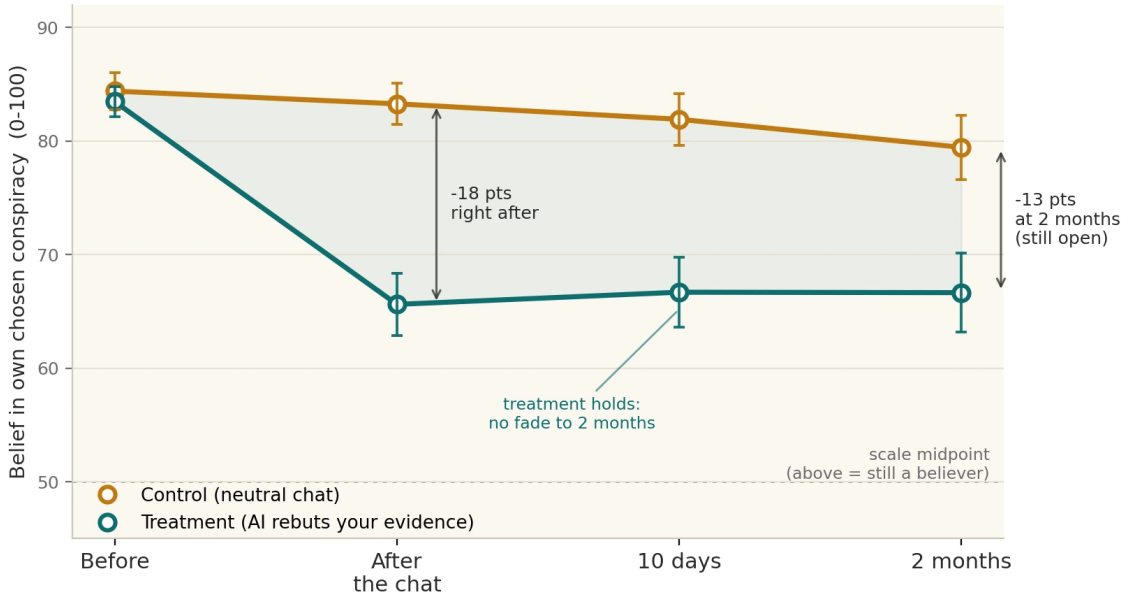
編輯部關切聲明是期刊附在一篇論文上的一則正式標記，以提醒讀者在問題仍處於釐清期間曾有疑問被提出。它**不是**撤稿（論文未被撤回），它本身也不是一項不當行為的發現。它意味著：帶著這一警告閱讀，因為記錄可能會改變。在此，在被告知這些問題後，作者進行了調查，向《科學》報告了具體情況，並提交了一份經修正的分析。

被標記的問題是具體的。作者被告知了參與者篩選標準在稿件與已發布的分析代碼之間的應用不一致，以及公共資料集因代碼合併錯誤而被拼接了無關行——這些問題使得部分報告的數字難以從已發布的材料中復現。作者向《科學》提交了一份經修正的分析管道和一組更新結果，他們表示這些結果在**方向、統計顯著性和實際效應量**上與原始結果一致。《科學》正在對其進行評估。在那項評估完成之前，精確的數字懸在一個只有期刊才能移除的問號之下。

所以這同時是兩個故事：一個關於事實是否能動搖根深蒂固的信念的有趣結果，以及一個科學記錄在公開中進行自我修正的鮮活案例。本文的要點是將它們區分開來。

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.

在這項研究中，據報告，一次反駁參與者自己陳述的證據的三輪 AI 對話，平均將該人選定陰謀論的信念削減了約 20%；與關於無關話題的對照聊天不同，這一下降在 10 天和 2 個月時仍然可測。報告的效果是持久的但部分的：大多數參與者之後仍然相信該陰謀論——是減少，而非治癒。該論文目前因報告不一致和公共資料集中的一個錯誤而處於《科學》（2026 年 6 月）的一項編輯部關切聲明之下；作者表示一項經修正的分析保留了效應的方向、顯著性和幅度，而《科學》仍在繼續評估。本圖表是由 The Clean Paper 從該公共資料集中重建的，因此所示數值是臨時的，並非論文的最終數字。

作者做了什麼

- 運行了兩項實驗，共 2190 名參與者，每人都用自己的話說出了一項他們相信的陰謀論，以及他們認為支持它的證據。
- 讓每位參與者與 GPT-4 Turbo 進行三輪對話。在**干預**條件下，AI 被提示去反駁參與者的具體證據；在**對照**條件下，它討論一個無關話題。
- 在對話前和緊接其後測量了對選定陰謀論的信念，然後在 **10 天和 2 個月**時重新聯繫了參與者。
- 讓一名專業事實核查員評估了 AI 在樣本中所做的全部 128 條事實性聲明的準確性。
- 檢查了效果是否溢出到其他無關聯的陰謀論——並且，作為特異性檢驗，是否也壓低了恰好是**真實**的陰謀論的信念。

他們發現了什麼

- **干預組相對於對照組，選定陰謀論的信念平均約 20% 的減少**（研究 1：95% 信賴區間 [13.8, 19.7] 個百分點於 100 分量表，$P < 0.001$）。
- **它是持久的**。在干預組中，下降並未消退：信念在 10 天和兩個月時保持在接近剛聊完時的水準而沒有回升，而對照組在此期間始終較高——論文將該效應描述為在選定陰謀論上至少持續兩個月未減弱，而非短暫搖擺。
- **它泛化了**，覆蓋了人們說出的各類陰謀論，甚至對初始信念很強的參與者也成立。
- **AI 的聲明在此設定中是準確的**：128 條中 127 條（99.2%）被評為真實，1 條（0.8%）誤導，0 條虛假。
- **它是特異的**，而非泛泛的懷疑：干預**沒有**降低對真實陰謀論的信念，表明它動搖了無支撐的信念，而非讓人對一切都變得犬儒。
- **它有適度的溢出**——對其他無關聯陰謀論的信念下降了約 12%，參與者報告了更強地反駁陰謀論主張的意願。主效應在研究 2 中得到維持，並在一個沒有對照組的小樣本中指向了同一方向（證據較弱，但一致）。

這沒有證明什麼

- 它**並非**目前不受質疑。該論文因資料處理和可復現性問題處於編輯部關切聲明之下，經修正的數字仍在《科學》的評估中。在審查結果出來前，將具體數值視為臨時性的。
- 它**沒有**表明 AI 能「治癒」陰謀論信念。約 20% 的平均減少是一次有意義的轉變，而非根除——大多數參與者之後仍然相信該理論，只是程度減輕。
- 這**不是**一次真實世界部署的結果。參與者自願加入研究並以誠意與 AI 互動；在現實中，對陰謀論最投入的人是最不可能去找一個會與他們爭辯的聊天機器人的。
- 99.2% 的準確率**特定於這一任務、這一模型和審慎的提示詞編寫**——並非語言模型陳述真實事物的普遍保證。作者明確指出，同樣的個人化說服能力如果被用於相反方向，也可能推動虛假信念。
- 「持久」在此處意味著**兩個月**，這對一次對話來說令人印象深刻，但與永久不是一回事。

證據有多強？

- **設計是一項真正的優勢**。受控的干預對對照實驗，乾淨的類安慰劑對照，跨兩項研究和一項獨立樣本的復現，持久性追蹤，以及對 AI 自身聲明準確性的明確檢查——這比大多數說服研究更為仔細，並且效應的方向在其被測試的每個地方都是一致的。
- **但編輯部關切聲明並非腳註**。能夠從已發布的資料和代碼中重新生成報告的數字是使一項結果值得信任的部分原因，而恰恰是這一點出了問題——篩選標準的不一致和來自代碼合併錯誤的重複行。作者的聲明——經修正的管道保留了結果——是令人鼓舞且合理的，但這是作者自己的說法，尚非獨立裁決。《科學》的評估是需要等待的東西。
- **誠實的地位是「被標記」，而非「被駁斥」或「被確認」**。一項構建良好、引人注目的研究，其確切數字正在接受正式審查。這是一個令人不舒服的地方來擱置一個好故事，且這正是準確的位置。

為什麼重要

多年來，一種有影響力的觀點認為，陰謀論信念是由心理需求和身份驅動，因此無法用事實與人爭論出來。一項持久的、基於證據的削減——如果能夠成立——是對這種悲觀情緒的有意義的反擊：它表明問題部分在於通用的駁斥從未觸達信徒實際引用的具體證據，而大語言模型可以大規模做到這一點。

它作為科學應如何運作的案例研究也同樣重要。編輯部關切聲明的教訓不是著名的結果一文不值；而是錯誤被浮現，資料被重新檢驗，聲明被公開地重新檢查。那個正在運行的機器是一個特徵，而非醜聞——這也正是對這一結果的正確姿態應是耐心而非掌聲或否定的原因。

而且它在 AI 上也有兩面性。用一個量身定製的論據來打消一個虛假信念的同樣能力，也能同樣有效地膨脹一個。這項技術是中性的；只有目的是不中性的。

清潔摘要

一項 2024 年的研究發現，與 GPT-4 Turbo 的三輪對話使人對其所持的一項陰謀論的信念降低了約 20%，該效應持續了兩個月並在各陰謀論類型間泛化，AI 的事實性聲明被評定為壓倒性準確。2026 年 6 月，該論文因資料處理和可復現性問題被置於一項編輯部關切聲明之下；作者報告稱一項經修正的分析在方向、顯著性和效應量上保留了該發現，《科學》仍在評估中。將其讀作一項真正有趣的、構建仔細的、目前正在審查中的結果——不是已解決的，不是已被駁斥的。關於它最誠實的一句話，正是過程本身正在書寫的那一句：再檢查一遍。

No-BS 檢查

論文所展示的：在受控實驗中，一次針對個人化證據的 AI 對話與選定陰謀論信念約 20% 的平均削減相關，且削減在兩個月內持續。**合理但未經證實的：**這一效應會在真實世界中復現；同樣的技術無法同樣好地用來灌輸虛假信念；經過修正的資料將保持不變。**它沒有展示的：**一個已完全復現、不受質疑的結果（目前）；AI 能根除陰謀論信念；人們會主動尋找此類對話。**主要侷限：**編輯部關切聲明正在審議中；實驗室設定而非真實世界；參與者自願且誠信互動；僅兩個月追蹤。**普通讀者應該抱有多大信心？**中等信心認為效應是真實的且方向正確，但具體數字是臨時的。低信心認為這種干預能夠在真實世界中復現。恰當的立場：一項真正有趣的研究，正在即時經歷科學自我修正的審查——保持耐心，不要預判。

來源

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