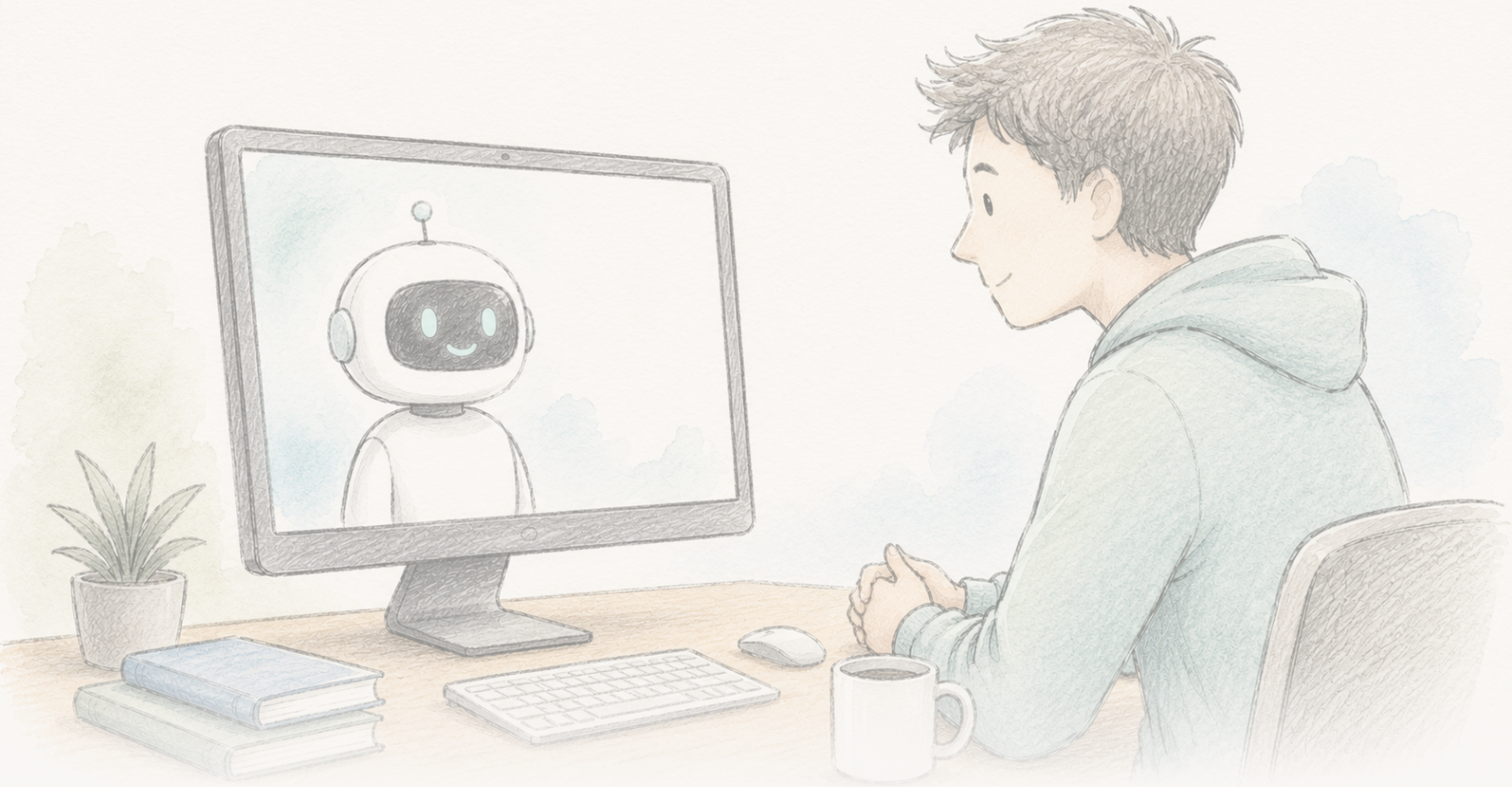


The Clean Paper

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

一次聊天機器人對話，似乎讓陰謀論信念降低了數月

一項 *2024* 年研究發現，與 *GPT-4 Turbo* 進行三輪對話，可讓人們對自己原本相信的某個陰謀論平均降低約 *20%* 的信念；效果持續兩個月、跨不同類型陰謀論都出現，而且 *AI* 提出的主張幾乎全被事實查核者判為正確。*2026* 年 *6* 月，論文因資料處理與可重現性問題被加註 *Editorial Expression of Concern*；作者表示更正後的分析在方向、統計顯著性與效應大小上都保留原本結果，而《*Science*》仍在評估。這是一項真正有意思、設計也相當仔細的結果——目前正在審查中，既未定案，也沒有被推翻。

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》已在這篇論文上附加 Editorial Expression of Concern，因為期刊正在評估資料處理與可重現性問題。這不是撤稿，也不是對研究不當行為的認定；它表示在審查完成前，報告的結果應視為暫定。

Editorial Expression of Concern →

事實畢竟可能有用——但論文現在掛著一面警示旗

2024 年，一項研究報告了一個違反常見直覺的結果。讓一個人和 AI 進行一段很短、只有三輪的對話——使用 GPT-4 Turbo，專門回應對方親自提出、用來支持自己所信陰謀論的證據——平均而言，那個人對該陰謀論的信念會下降約 20%。兩個月後，這個下降仍然存在。效果出現在經典陰謀論（John F. Kennedy 遇刺、外星人、光明會）以及較新的議題（COVID-19、2020 年美國大選），甚至對原本信念很強、而且把它與自身認同綁在一起的人也有效。專業事實查核者評估 AI 在樣本中提出的 128 項事實主張後，127 項被判為真、1 項具誤導性，沒有任何一項被判為假。

最廣為流傳的說法是：你真的可能把人從陰謀論兔子洞裡勸出來——相信陰謀論的人並不是事實無法觸及，只是需要足夠具體、正好對準他們所引用證據的事實。這是一個真正有意思的主張，而且這項研究的設計比大多數說服研究更仔細。

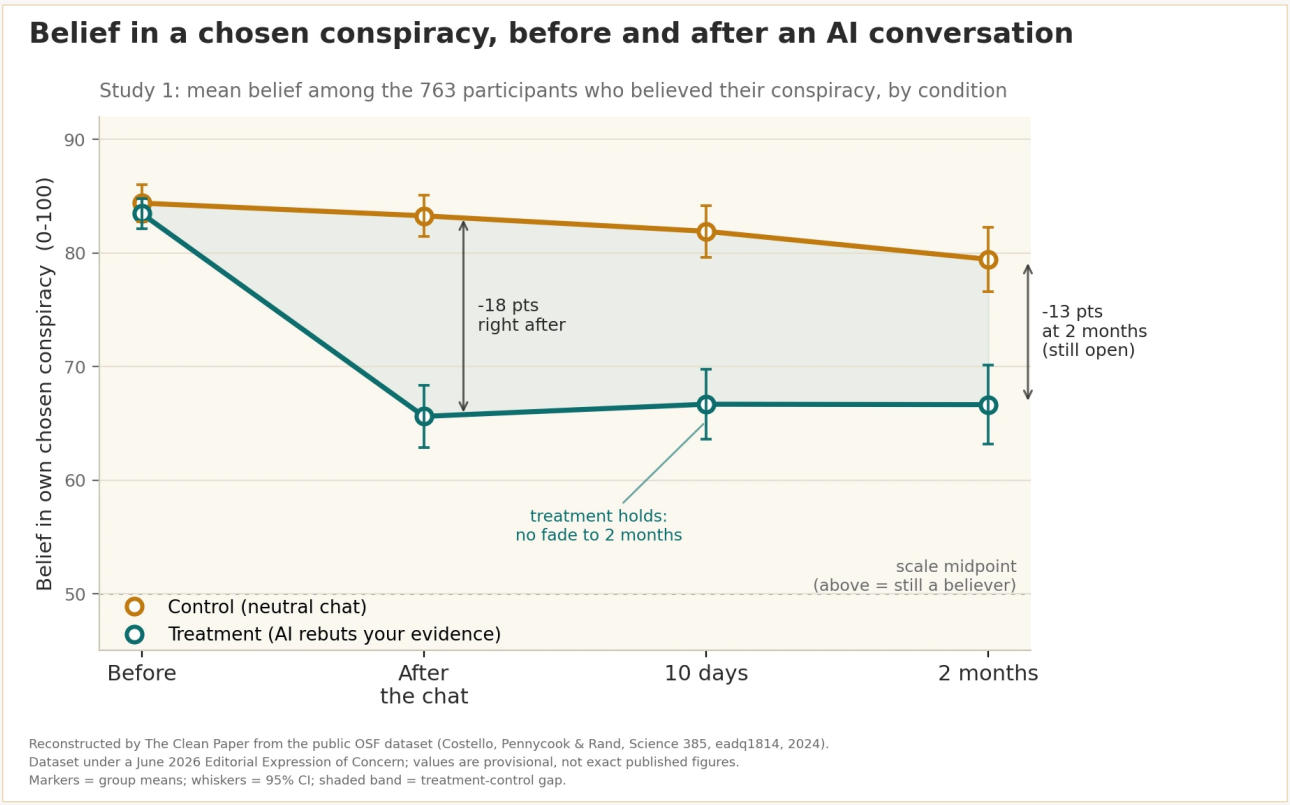
接著到了 2026 年 6 月，《Science》在論文上加了一則 **Editorial Expression of Concern**。這一段是很多轉述最可能略過的，但它恰恰決定了目前我們可以把多大重量放在這項結果上。

Editorial Expression of Concern 是什麼——又不是什麼

Editorial Expression of Concern 是期刊加在論文上的正式警示，表示有人提出了需要處理的問題，而調查或釐清仍在進行中。它**不是**撤稿（論文並未被撤回），本身也不是研究不當行為的認定。它的意思是：閱讀時必須把這個警示放在心上，因為正式紀錄之後可能改變。在這個案例裡，作者得知問題後自行調查，向《Science》報告細節，並提交了一套更正後的分析。

被指出的問題相當具體。作者後來得知，論文文字與已公開的分析程式碼，在受試者篩選條件的實際套用方式上存在不一致；同時，公開資料集中還混入了由程式碼合併錯誤拼接進去的額外列。這些問題使得部分報告數字無法順利從公開材料重現。作者已提供給《Science》一套更正後的分析流程與更新結果，並表示新結果在**方向、統計顯著性與實質效應大小**上都與原始結果一致。《Science》仍在評估。在期刊完成評估之前，精確數字都應該被視為帶著問號的暫定值，而只有期刊能在最後把這個問號拿掉。

所以這其實是兩個故事同時發生：一個關於事實能否動搖根深蒂固信念的有趣結果，以及一個科學紀錄如何在公開環境中自我修正的即時案例。這篇文章要做的，就是不要把兩件事混在一起。



研究原本報告：針對受試者自己提出的證據，由 AI 進行三輪反駁式對話，可讓他們對所選陰謀論的信念平均下降約 20%；不同於討論無關主題的控制組，這個下降在 10 天與 2 個月後仍可量測到。效果持久但有限：多數人之後仍然相信該陰謀論，只是程度降低——這是減少，不是「治癒」。這篇論文目前因報告不一致與公開資料集中的錯誤，自 2026 年 6 月起被《Science》加註 Editorial Expression of Concern；作者表示更正後分析保留效應的方向、顯著性與大小，而《Science》仍在評估。這張圖由 The Clean Paper 依公開資料集重新繪製，因此其中數值應視為暫定，而不是論文的最終版本。

作者做了什麼

- 進行兩項實驗，共納入 2,190 名參與者；每個人都用自己的話說出一個自己相信的陰謀論，以及他們認為支持它的證據。
- 每位參與者與 GPT-4 Turbo 進行三輪對話。在**處理組**中，AI 被要求直接反駁參與者提出的特定證據；在**控制組**中，AI 則討論一個無關主題。
- 在對話前、對話後立即測量對所選陰謀論的信念，並於 **10 天後與 2 個月後**再次聯繫參與者。
- 請專業事實查核者評估樣本中 AI 提出的全部 128 項事實主張是否正確。
- 檢查效果是否會外溢到其他、不相關的陰謀論；並作為特異性測試，看看它是否也會讓人降低對那些碰巧**真的發生過**的陰謀事件的相信程度。

他們發現了什麼

- 與控制組相比，處理組對自己所選陰謀論的信念**平均下降約 20%**（研究 1：100 分量表上的 95% 信賴區間為 [13.8, 19.7] 分，$P < 0.001$）。
- **效果持續存在**。處理組的下降沒有明顯反彈：10 天與兩個月後，信念都維持在接近對話剛結束時的水準，而控制組全程較高。論文將這描述為對所選陰謀論的影響至少持續兩個月，並未減弱，而不是短暫波動。
- **效果具有廣泛性**。不論人們提出哪一類陰謀論，整體方向都類似；即使原本信念很強的人，也仍出現效果。
- **在這個任務中，AI 的事實主張高度準確**。 128 項中有 127 項（99.2%）被判為真、1 項（0.8%）具誤導性、0 項為假。
- **效果具有特異性，而不是讓人全面懷疑**。介入並沒有降低參與者對真實陰謀事件的相信程度，表示它更像是在削弱缺乏支持的信念，而不是把人變得對一切都犬儒。
- **也有一些外溢效果**。對其他不相關陰謀論的信念約下降 12%，參與者也更表示願意反駁陰謀論主張。主要效果在研究 2 中重現；另一個較小、沒有控制組的樣本也朝相同方向變化（證據較弱，但一致）。

這並沒有證明什麼

- 目前它**不是一個毫無爭議的結果**。論文因資料處理與可重現性問題被加註 Editorial Expression of Concern，而且更正後數字仍在《Science》評估中。在審查完成前，具體數值都應視為暫定。
- 它**沒有**證明 AI 能「治好」陰謀論信念。平均下降約 20% 是有意義的變化，但不是把信念消除——多數人之後仍相信，只是沒有原來那麼強。
- 這**不是**真實世界部署的結果。參與者自願加入研究，並願意認真與 AI 互動；在現實世界，最投入某個陰謀論的人，往往也是最不可能主動找一個會和自己辯論的聊天機器人的人。
- 99.2% 的準確率**只適用於這個任務、這個模型與這套仔細設計的提示**，不能當成大型語言模型通常都會說真話的保證。作者也明確指出，同樣的個人化說服能力，如果拿去推動錯誤信念，也可能一樣有效。
- 這裡所謂「持久」是指**兩個月**。對一次對話來說已經相當可觀，但並不等於永久。

證據有多強

- **研究設計本身是明顯優點**。有處理組與控制組的實驗、設計清楚的安慰劑式控制、兩項研究與另一個獨立樣本、長期追蹤，以及對 AI 自己說法的準確度檢查——這比多數說服研究嚴謹，而且只要做了測試，效應方向大致都一致。
- **但 Expression of Concern 絕不是腳註**。能否從公開資料與程式碼重新產生報告數字，本來就是結果可信度的一部分；而現在出問題的恰恰就是這裡——篩選條件的不一致，以及程式碼合併錯誤造成的重複／額外資料列。作者表示更正後流程仍保留結果，這是令人鼓舞、也合理的訊息，但目前仍是作者自己的說法，還不是獨立裁定。《Science》的評估結果才是關鍵。
- **最誠實的狀態是「被正式標記」，不是「被推翻」或「已確認」**。這是一項設計仔細、結果吸引人的

研究，但其精確數字正接受正式審查。這不是一個好故事最舒服的結尾，卻是現在最準確的位置。

為什麼重要

多年來，一種頗具影響力的看法認為，陰謀論信念主要由心理需求與身分認同驅動，因此不可能靠事實辯論把人說服。如果這個持久、以證據為核心的降低效果最後能站穩，它會對這種悲觀論形成重要制衡：也許問題的一部分在於，一般化的闢謠從未真正對準信徒自己拿來支持信念的特定證據，而大型語言模型剛好可以大規模做到這件事。

它也重要，因為這是科學應如何運作的一個案例。Expression of Concern 的教訓不是「轟動研究都不值得相信」，而是錯誤會被指出、資料會被重新檢查、主張會在公開環境中再次被驗證。這套機制正在運轉，是功能，不是醜聞——也正因如此，對這項結果最合理的態度是耐心，而不是立刻歡呼或立刻否定。

而且這對 AI 是雙面刃。同一種能用量身打造的論點削弱錯誤信念的能力，也可能同樣有效地強化錯誤信念。技術本身是中性的，目的不是。

重點摘要

一項 2024 年研究發現，與 GPT-4 Turbo 進行三輪對話，可讓人們對自己所信某個陰謀論的信念平均降低約 20%；效果持續兩個月、跨不同陰謀論類型都出現，而 AI 的事實主張幾乎全被評為正確。2026 年 6 月，論文因資料處理與可重現性問題被加註 Editorial Expression of Concern；作者表示更正後分析在方向、顯著性與大小上仍保留原本結果，《Science》則仍在評估。現在最恰當的讀法，是把它視為一項真正有意思、設計仔細、但正在正式審查中的結果——還沒有定案，也沒有被推翻。對它最誠實的一句話，正是科學流程本身正在寫的那句：再檢查一次。

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