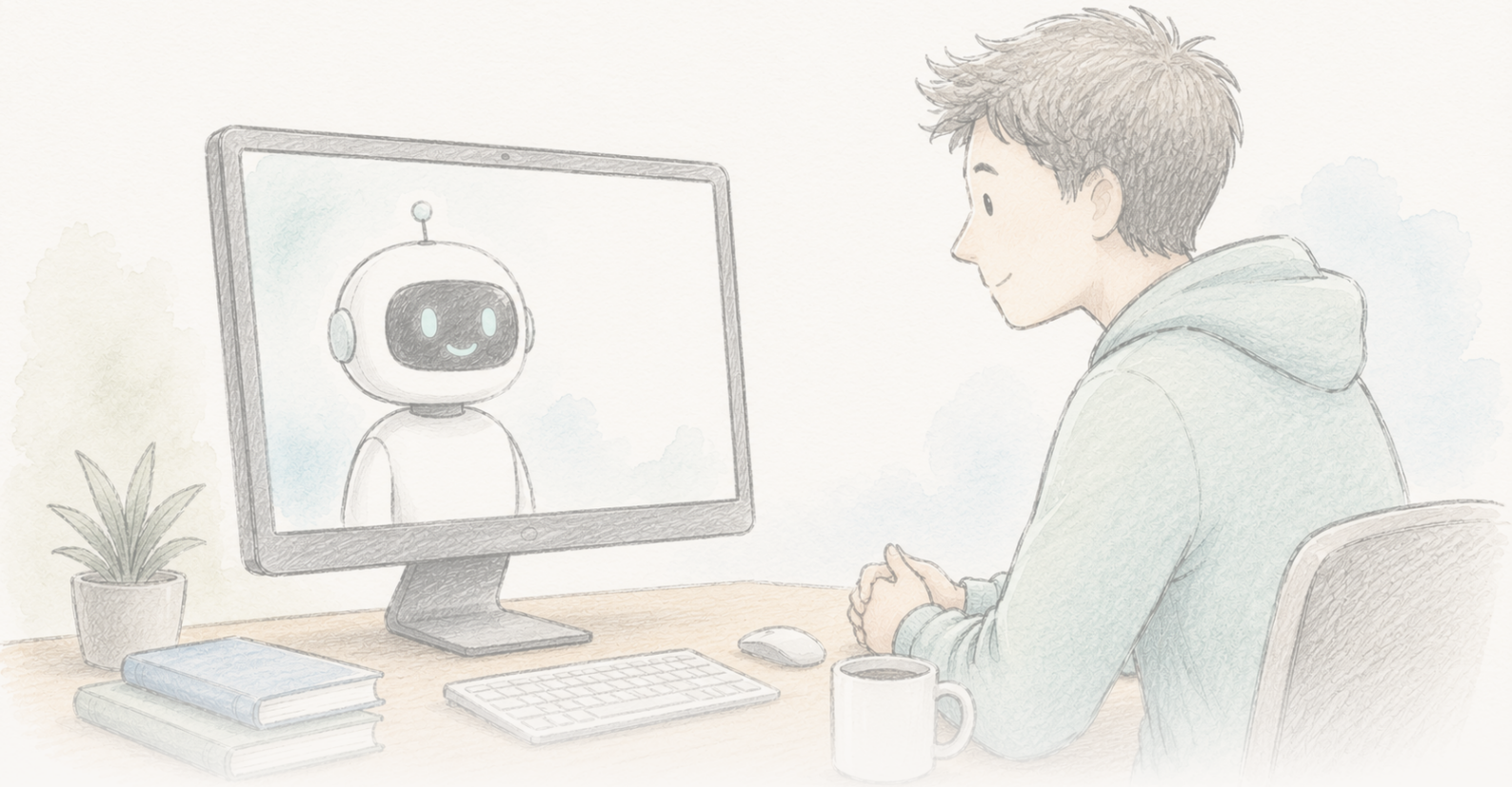


The Clean Paper

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

一次聊天机器人对话，似乎让阴谋论信念降低了数月

一项 *2024* 年研究发现，与 *GPT-4 Turbo* 进行三轮对话，可让人们对自己原本相信的某个阴谋论平均降低约 *20%* 的信念；效果持续两个月、跨不同类型阴谋论都出现，而且 *AI* 提出的主张几乎全被事实核查人员判为正确。*2026* 年 *6* 月，论文因数据处理与可复现性问题被加注 *Editorial Expression of Concern*；作者表示更正后的分析在方向、统计显著性与效应量上都保留原本结果，而《*Science*》仍在评估。这是一项真正有意思、设计也相当仔细的结果——目前正在审查中，既无最终定论，也没有被推翻。

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》已在这篇论文上附加 Editorial Expression of Concern，因为期刊正在评估数据处理与可复现性问题。这不是撤稿，也不是对研究不当行为的认定；它表示在审查完成前，报告的结果应视为暂定。

Editorial Expression of Concern →

事实毕竟可能有用——但论文现在挂着一面警示旗

2024 年，一项研究报告了一个违反常见直觉的结果。让一个人和 AI 进行一段很短、只有三轮的对话——使用 GPT-4 Turbo，专门回应对方亲自提出、用来支持自己所信阴谋论的证据——平均而言，那个人对该阴谋论的信念会下降约 20%。两个月后，这个下降仍然存在。效果出现在经典阴谋论（John F. Kennedy 遇刺、外星人、光明会）以及较新的议题（COVID-19、2020 年美国大选），甚至对原本信念很强、而且把它与自身认同绑在一起的人也有效。专业事实核查人员评估 AI 在样本中提出的 128 项事实主张后，127 项被判为真、1 项具误导性，没有任何一项被判为假。

最广为流传的说法是：你真的可能把人从阴谋论兔子洞里劝出来——相信阴谋论的人并不是事实无法触及，只是需要足够具体、正好对准他们所引用证据的事实。这是一个真正有意思的主张，而且这项研究的设计比大多数说服研究更仔细。

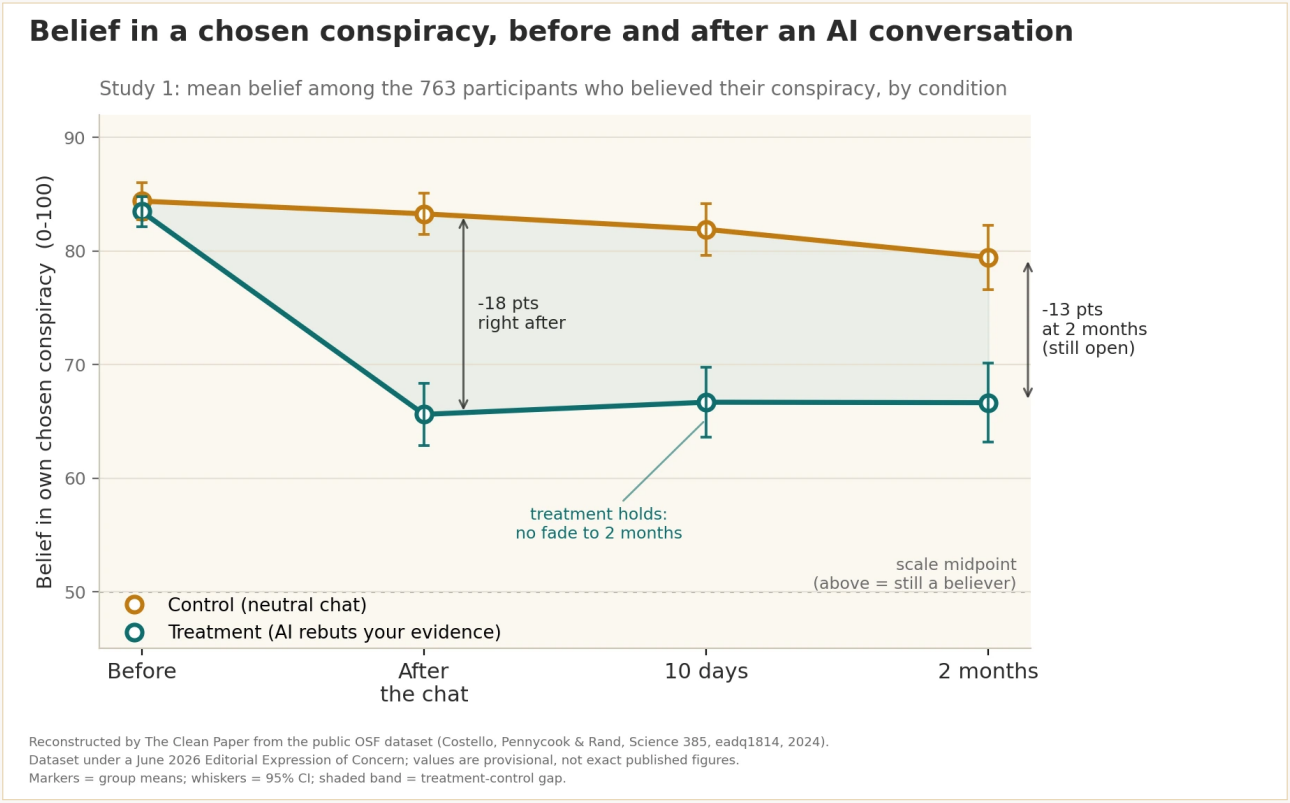
接着到了 2026 年 6 月，《Science》在论文上加了一则 **Editorial Expression of Concern**。这一段是很多转述最可能略过的，但它恰恰决定了目前我们可以把多大重量放在这项结果上。

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

Editorial Expression of Concern 是期刊加在论文上的正式警示，表示有人提出了需要处理的问题，而调查或澄清仍在进行中。它**不是**撤稿（论文并未被撤回），本身也不是研究不当行为的认定。它的意思是：阅读时必须把这个警示放在心上，因为正式记录之后可能改变。在这个案例里，作者得知问题后自行调查，向《Science》报告细节，并提交了一套更正后的分析。

被指出的问题相当具体。作者后来得知，论文文字与已公开的分析代码，在受试者筛选条件的实际套用方式上存在不一致；同时，公开数据集中还混入了由代码合并错误拼接进去的额外列。这些问题使得部分报告数字无法顺利从公开材料重现。作者已提供给《Science》一套更正后的分析流程与更新结果，并表示新结果在**方向、统计显著性与效应量**上都与原始结果一致。《Science》仍在评估。在期刊完成评估之前，精确数字都应该被视为带着问号的暂定值，而只有期刊能在最后把这个问号拿掉。

所以这其实是两个故事同时发生：一个关于事实能否动摇根深蒂固信念的有趣结果，以及一个科学记录如何在公开环境中自我修正的实时案例。这篇文章要做的，就是不要把两件事混在一起。



研究原本报告：针对受试者自己提出的证据，由 AI 进行三轮反驳式对话，可让他们对所选阴谋论的信念平均下降约 20%；不同于讨论无关主题的控制组，这个下降在 10 天与 2 个月后仍可测得。效果持久但有限：多数人之后仍然相信该阴谋论，只是程度降低——这是减少，不是“治愈”。这篇论文目前因报告不一致与公开数据集中的错误，自 2026 年 6 月起被《Science》加注 Editorial Expression of Concern；作者表示更正后分析保留效应的方向、显著性与大小，而《Science》仍在评估。这张图由 The Clean Paper 依公开数据集重新绘制，因此其中数值应视为暂定，而不是论文的最终版本。

作者做了什么

- 进行两项实验，共纳入 2,190 名参与者；每个人都用自己的话说出一个自己相信的阴谋论，以及他们认为支持它的证据。
- 每位参与者与 GPT-4 Turbo 进行三轮对话。在**处理组**中，AI 被要求直接反驳参与者提出的特定证据；在**控制组**中，AI 则讨论一个无关主题。
- 在对话前、对话后立即测量对所选阴谋论的信念，并于 **10 天后与 2 个月后**再次联系参与者。
- 请专业事实核查人员评估样本中 AI 提出的全部 128 项事实主张是否正确。
- 检查效果是否会溢出到其他、不相关的阴谋论；并作为特异性测试，看看它是否也会让人降低对那些碰巧**真的发生过**的阴谋事件的相信程度。

他们发现了什么

- 与控制组相比，处理组对自己所选阴谋论的信念**平均下降约 20%**（研究 1：100 分量表上的 95% 置信区间为 [13.8, 19.7] 分，$P < 0.001$）。
- **效果持续存在**。处理组的下降没有明显反弹：10 天与两个月后，信念都维持在接近对话刚结束时的水平，而控制组全程较高。论文将这描述为对所选阴谋论的影响至少持续两个月，并未减弱，而不是短暂波动。
- **效果具有广泛性**。不论人们提出哪一类阴谋论，整体方向都类似；即使原本信念很强的人，也仍出现效果。
- **在这个任务中，AI 的事实主张高度准确**。128 项中有 127 项（99.2%）被判为真、1 项（0.8%）具误导性、0 项为假。
- **效果具有特异性，而不是让人全面怀疑**。干预并没有降低参与者对真实阴谋事件的相信程度，表示它更像是在削弱缺乏支持的信念，而不是把人变得对一切都犬儒。
- **也有一些溢出效果**。对其他不相关阴谋论的信念约下降 12%，参与者也更表示愿意反驳阴谋论主张。主要效果在研究 2 中重现；另一个较小、没有控制组的样本也朝相同方向变化（证据较弱，但一致）。

这并没有证明什么

- 目前它**不是一个毫无争议的结果**。论文因数据处理与可复现性问题被加注 Editorial Expression of Concern，而且更正后数字仍在《Science》评估中。在审查完成前，具体数值都应视为暂定。
- 它**没有**证明 AI 能“治好”阴谋论信念。平均下降约 20% 是有意义的变化，但不是把信念消除——多数人之后仍相信，只是没有原来那么强。
- 这**不是**真实世界部署的结果。参与者自愿加入研究，并愿意认真与 AI 互动；在现实世界，最投入某个阴谋论的人，往往也是最不可能主动找一个会和自己辩论的聊天机器人的人。
- 99.2% 的准确率**只适用于这个任务、这个模型与这套仔细设计的提示**，不能当成大型语言模型通常都会说真话的保证。作者也明确指出，同样的个性化说服能力，如果拿去推动错误信念，也可能一样有效。
- 这里所谓“持久”是指**两个月**。对一次对话来说已经相当可观，但并不等于永久。

证据有多强

- **研究设计本身是明显优点**。有处理组与控制组的实验、设计清楚的安慰剂式控制、两项研究与另一个独立样本、长期追踪，以及对 AI 自己说法的准确度检查——这比多数说服研究严谨，而且只要做了测试，效应方向大致都一致。
- **但 Expression of Concern 绝不是脚注**。能否从公开数据与代码重新产生报告数字，本来就是结果可信度的一部分；而现在出问题的恰恰就是这里——筛选条件的不一致，以及代码合并错误造成的重复/额外数据列。作者表示更正后流程仍保留结果，这是令人鼓舞、也合理的信息，但目前仍是作者自己的说法，还不是独立裁定。《Science》的评估结果才是关键。
- **最诚实的状态是“被正式标记”，不是“被推翻”或“已确认”**。这是一项设计仔细、结果吸引人的研

究，但其精确数字正接受正式审查。这不是一个好故事最舒服的结尾，却是现在最准确的位置。

为什么重要

多年来，一种颇具影响力的看法认为，阴谋论信念主要由心理需求与身份认同驱动，因此不可能靠事实辩论把人说服。如果这个持久、以证据为核心的降低效果最后能站稳，它会对这种悲观论形成重要制衡：也许问题的一部分在于，一般化的辟谣从未真正对准信徒自己拿来支持信念的特定证据，而大型语言模型刚好可以大规模做到这件事。

它也重要，因为这是科学应如何运作的一个案例。Expression of Concern 的教训不是“轰动研究都不值得相信”，而是错误会被指出、数据会被重新检查、主张会在公开环境中再次被验证。这套机制正在运转，是功能，不是丑闻——也正因如此，对这项结果最合理的态度是耐心，而不是立刻欢呼或立刻否定。

而且这对 AI 是双面刃。同一种能用量身打造的论点削弱错误信念的能力，也可能同样有效地强化错误信念。技术本身是中性的，目的不是。

简明总结

一项 2024 年研究发现，与 GPT-4 Turbo 进行三轮对话，可让人们对自己所信某个阴谋论的信念平均降低约 20%；效果持续两个月、跨不同阴谋论类型都出现，而 AI 的事实主张几乎全被评为正确。2026 年 6 月，论文因数据处理与可复现性问题被加注 Editorial Expression of Concern；作者表示更正后分析在方向、显著性与大小上仍保留原本结果，《Science》则仍在评估。现在最恰当的读法，是把它视为一项真正有意思、设计仔细、但正在正式审查中的结果——尚无定论，也没有被推翻。对它最诚实的一句话，正是科学流程本身正在写的那句：再检查一次。

来源

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