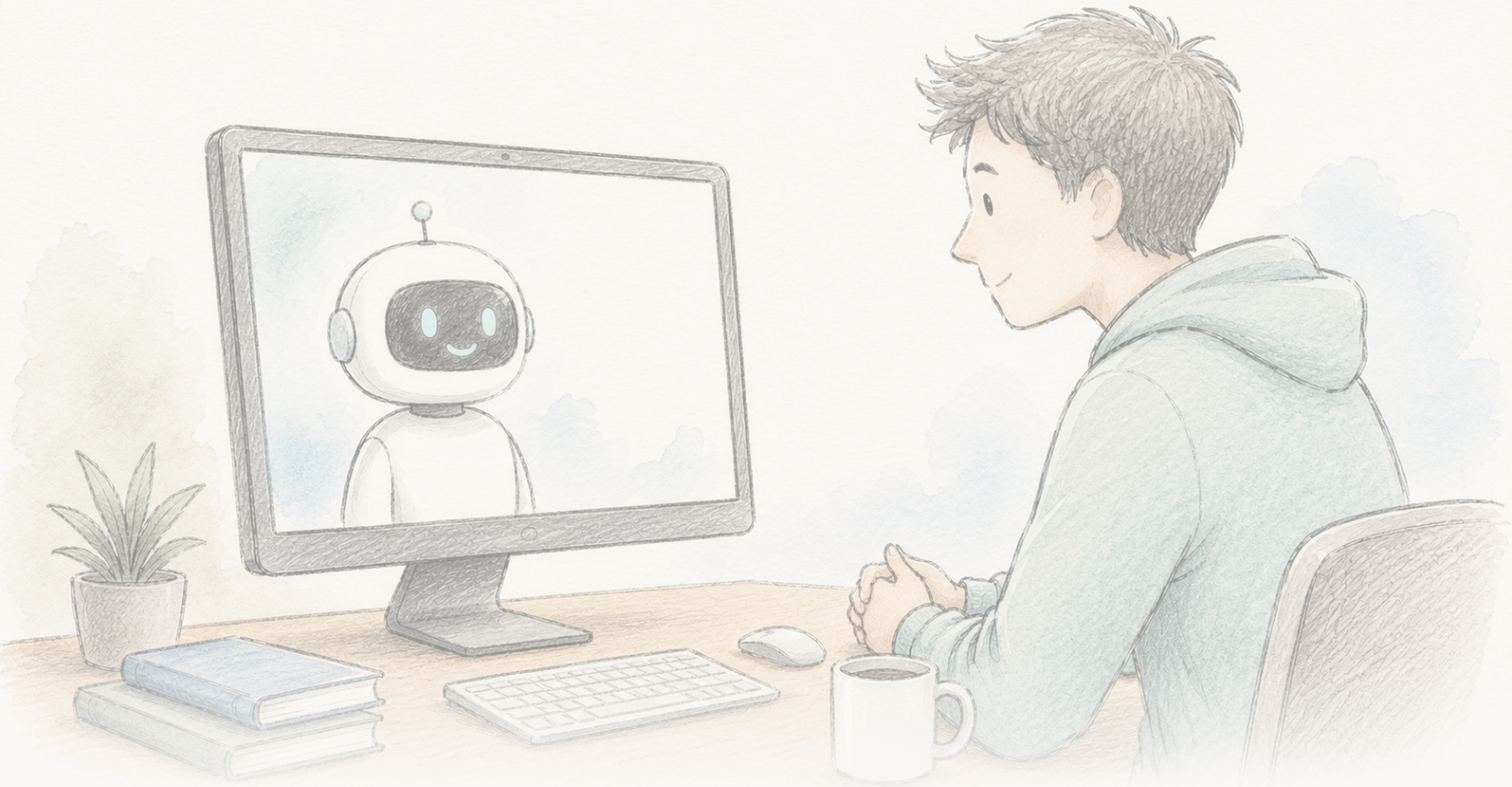


The Clean Paper

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

一个聊天机器人似乎持续减少了阴谋论信念达数月

一项 *2024* 年的研究发现，与 *GPT-4 Turbo* 的三轮对话使人对其所持的一项阴谋论的信念削减了约 *20%*，该效应持续了两个月，并在各阴谋论类型间泛化，且基于被评定为压倒性准确的 *AI* 声明。*2026* 年 *6* 月，该论文因数据处理和可复现性问题被置于一项编辑部关切声明之下；作者表示一项经修正的分析在方向、显著性和效应量上保留了该发现，《科学》仍在评估中。一项真正有趣的、构建仔细的——目前正在审查中的结果，既非已解决，也非已被驳斥。

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

《科学》在评估数据处理和可复现性问题期间，已对该论文附上一则编辑部关切声明。这不是撤稿，也不是不当行为的发现；它意味着报告的结果在其审查解决之前应被视为临时的。

Editorial Expression of Concern →

事实，终究——以及论文上的一面旗帜

2024 年，一项研究报告了一件与一条舒适的常识常理相悖的发现。让人与 AI（GPT-4 Turbo，被提示去反驳他们个人相信的一项阴谋论所引用的具体证据）进行一次简短的三轮对话，平均而言，他们对该理论的信念下降了约 20%。这一下降在两个月后仍然存在。它跨越了经典阴谋论（肯尼迪遇刺、外星人、光明会）和时政阴谋论（COVID-19、2020 年美国大选），甚至在那些信念强烈且与身份绑定的参与者中也有效。当一名专业事实核查员对 AI 所做的 128 条声明的样本进行评估时，127 条为真，1 条为误导，0 条为假。

广为流传的标题是，你可以把人从兔子洞里拉出来——阴谋论者并非超出证据的触及范围，他们只是需要正确的证据，以足够具体的方式传递。这是一个真正有趣的主张，该研究的构建比大多数说服研究更为仔细。

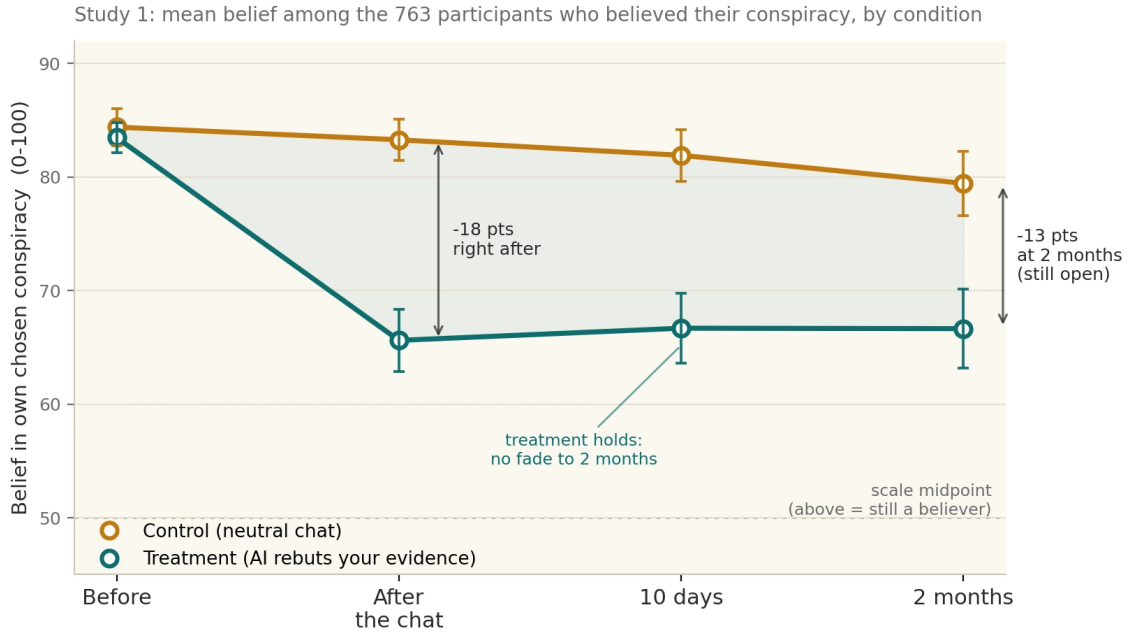
然后，在 2026 年 6 月，《科学》对该论文发布了一则**编辑部关切声明**。这是大多数复述会跳过的部分，而这恰恰是决定这一结果目前能承受多大分量的部分。

编辑部关切声明是什么——以及不是什么

编辑部关切声明是期刊附在一篇论文上的一则正式标记，以提醒读者在问题仍处于厘清期间曾有疑问被提出。它**不是**撤稿（论文未被撤回），它本身也不是一项不当行为的发现。它意味着：带着这一警告阅读，因为记录可能会改变。在此，在被告知这些问题后，作者进行了调查，向《科学》报告了具体情况，并提交了一份经修正的分析。

被标记的问题是具体的。作者被告知了参与者筛选标准在稿件与已发布的分析代码之间的应用不一致，以及公共数据集因代码合并错误而被拼接了无关行——这些问题使得部分报告的数字难以从已发布的材料中复现。作者向《科学》提交了一份经修正的分析管道和一组更新结果，他们表示这些结果在**方向、统计显著性和实际效应量**上与原始结果一致。《科学》正在对其进行评估。在那项评估完成之前，精确的数字悬在一个只有期刊才能移除的问号之下。

所以这同时是两个故事：一个关于事实是否能动摇根深蒂固的信念的有趣结果，以及一个科学记录在公开中进行自我修正的鲜活案例。本文的要点是将它们区分开来。



在这项研究中，据报告，一次反驳参与者自己陈述的证据的三轮 AI 对话，平均将该人选定阴谋论的信念削减了约 20%；与关于无关话题的对照聊天不同，这一下降在 10 天和 2 个月时仍然可测。报告的效果是持久的但部分的：大多数参与者之后仍然相信该阴谋论——是减少，而非治愈。该论文目前因报告不一致和公共数据集中的一个错误而处于《科学》（2026 年 6 月）的一项编辑部关切声明之下；作者表示一项经修正的分析保留了效应的方向、显著性和幅度，而《科学》仍在继续评估。本图表是由 The Clean Paper 从该公共数据集中重建的，因此所示数值是临时的，并非论文的最终数字。

作者做了什么

- 运行了两项实验，共 2190 名参与者，每人都用自己的话说出了一项他们相信的阴谋论，以及他们认为支持它的证据。
- 让每位参与者与 GPT-4 Turbo 进行三轮对话。在**干预**条件下，AI 被提示去反驳参与者的具体证据；在**对照**条件下，它讨论一个无关话题。
- 在对话前和紧接其后测量了对选定阴谋论的信念，然后在 **10 天和 2 个月**时重新联系了参与者。
- 让一名专业事实核查员评估了 AI 在样本中所做的全部 128 条事实性声明的准确性。
- 检查了效果是否溢出到其他无关联的阴谋论——并且，作为特异性检验，是否也压低了恰好是**真实**的阴谋论的信念。

他们发现了什么

- **干预组相对于对照组，选定阴谋论的信念平均约 20% 的减少**（研究 1：95% 置信区间 [13.8, 19.7] 个百分点于 100 分量表，$P < 0.001$）。
- **它是持久的**。在干预组中，下降并未消退：信念在 10 天和两个月时保持在接近刚聊完时的水平而没有回升，而对照组在此期间始终较高——论文将该效应描述为在选定阴谋论上至少持续两个月未减弱，而非短暂摇摆。
- **它泛化了**，覆盖了人们说出的各类阴谋论，甚至对初始信念很强的参与者也成立。
- **AI 的声明在此设置中是准确的**：128 条中 127 条（99.2%）被评为真实，1 条（0.8%）误导，0 条虚假。
- **它是特异的**，而非泛泛的怀疑：干预**没有**降低对真实阴谋论的信念，表明它动摇了无支撑的信念，而非让人对一切都变得犬儒。
- **它有适度的溢出**——对其他无关联阴谋论的信念下降了约 12%，参与者报告了更强地反驳阴谋论主张的意愿。主效应在研究 2 中得到维持，并在一个没有对照组的小样本中指向了同一方向（证据较弱，但一致）。

这没有证明什么

- 它**并非**目前不受质疑。该论文因数据处理和可复现性问题处于编辑部关切声明之下，经修正的数字仍在《科学》的评估中。在审查结果出来前，将具体数值视为临时性的。
- 它**没有**表明 AI 能”治愈”阴谋论信念。约 20% 的平均减少是一次有意义的转变，而非根除——大多数参与者之后仍然相信该理论，只是程度减轻。
- 这**不是**一次真实世界部署的结果。参与者自愿加入研究并以诚意与 AI 互动；在现实中，对阴谋论最投入的人是最不可能去找一个会与他们争辩的聊天机器人的。
- 99.2% 的准确率**特定于这一任务、这一模型和审慎的提示词编写**——并非语言模型陈述真实事物的普遍保证。作者明确指出，同样的个性化说服能力如果被用于相反方向，也可能推动虚假信念。
- “持久”在此处意味着**两个月**，这对一次对话来说令人印象深刻，但与永久不是一回事。

证据有多强?

- **设计是一项真正的优势**。受控的干预对对照实验，干净的类安慰剂对照，跨两项研究和一项独立样本的复现，持久性追踪，以及对 AI 自身声明准确性的明确检查——这比大多数说服研究更为仔细，并且效应的方向在其被测试的每个地方都是一致的。
- **但编辑部关切声明并非脚注**。能够从已发布的数据和代码中重新生成报告的数字是使一项结果值得信任的部分原因，而恰恰是这一点出了问题——筛选标准的不一致和来自代码合并错误的重复行。作者的声明——经修正的管道保留了结果——是令人鼓舞且合理的，但这是作者自己的说法，尚非独立裁决。《科学》的评估是需要等待的东西。
- **诚实的地位是”被标记”，而非”被驳斥”或”被确认”**。一项构建良好、引人注目的研究，其确切数字正在接受正式审查。这是一个令人不舒服的地方来搁置一个好故事，且这正是准确的位置。

为什么重要

多年来，一种有影响力的观点认为，阴谋论信念是由心理需求和身份驱动，因此无法用事实与人争论出来。一项持久的、基于证据的削减——如果能够成立——是对这种悲观情绪的有意义的反击：它表明问题部分在于通用的驳斥从未触达信徒实际引用的具体证据，而大语言模型可以大规模做到这一点。

它作为科学应如何运作的案例研究也同样重要。编辑部关切声明的教训不是著名的结果一文不值；而是错误被浮现，数据被重新检验，声明被公开地重新检查。那个正在运行的机器是一个特征，而非丑闻——这也正是对这一结果的正确姿态应是耐心而非掌声或否定的原因。

而且它在 AI 上也有两面性。用一个量身定制的论据来打消一个虚假信念的同样能力，也能同样有效地膨胀一个。这项技术是中性的；只有目的是不中性的。

清洁摘要

一项 2024 年的研究发现，与 GPT-4 Turbo 的三轮对话使人对其所持的一项阴谋论的信念降低了约 20%，该效应持续了两个月并在各阴谋论类型间泛化，AI 的事实性声明被评定为压倒性准确。2026 年 6 月，该论文因数据处理和可复现性问题被置于一项编辑部关切声明之下；作者报告称一项经修正的分析在方向、显著性和效应量上保留了该发现，《科学》仍在评估中。将其读作一项真正有趣的、构建仔细的、目前正在审查中的结果——不是已解决的，不是已被驳斥的。关于它最诚实的一句话，正是过程本身正在书写的那一句：再检查一遍。

No-BS 检查

论文所展示的：在受控实验中，一次针对个人化证据的 AI 对话与选定阴谋论信念约 20% 的平均削减相关，且削减在两个月内持续。**合理但未经证实的：**这一效应会在真实世界中复现；同样的技术无法同样好地用来灌输虚假信念；经过修正的数据将保持不变。**它没有展示的：**一个已完全复现、不受质疑的结果（目前）；AI 能根除阴谋论信念；人们会主动寻找此类对话。**主要局限：**编辑部关切声明正在审议中；实验室设置而非真实世界；参与者自愿且诚信互动；仅两个月追踪。**普通读者应该抱有多大信心？**中等信心认为效应是真实的且方向正确，但具体数字是临时的。低信心认为这种干预能够在真实世界中复现。恰当的立场：一项真正有趣的研究，正在实时经历科学自我修正的审查——保持耐心，不要预判。

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