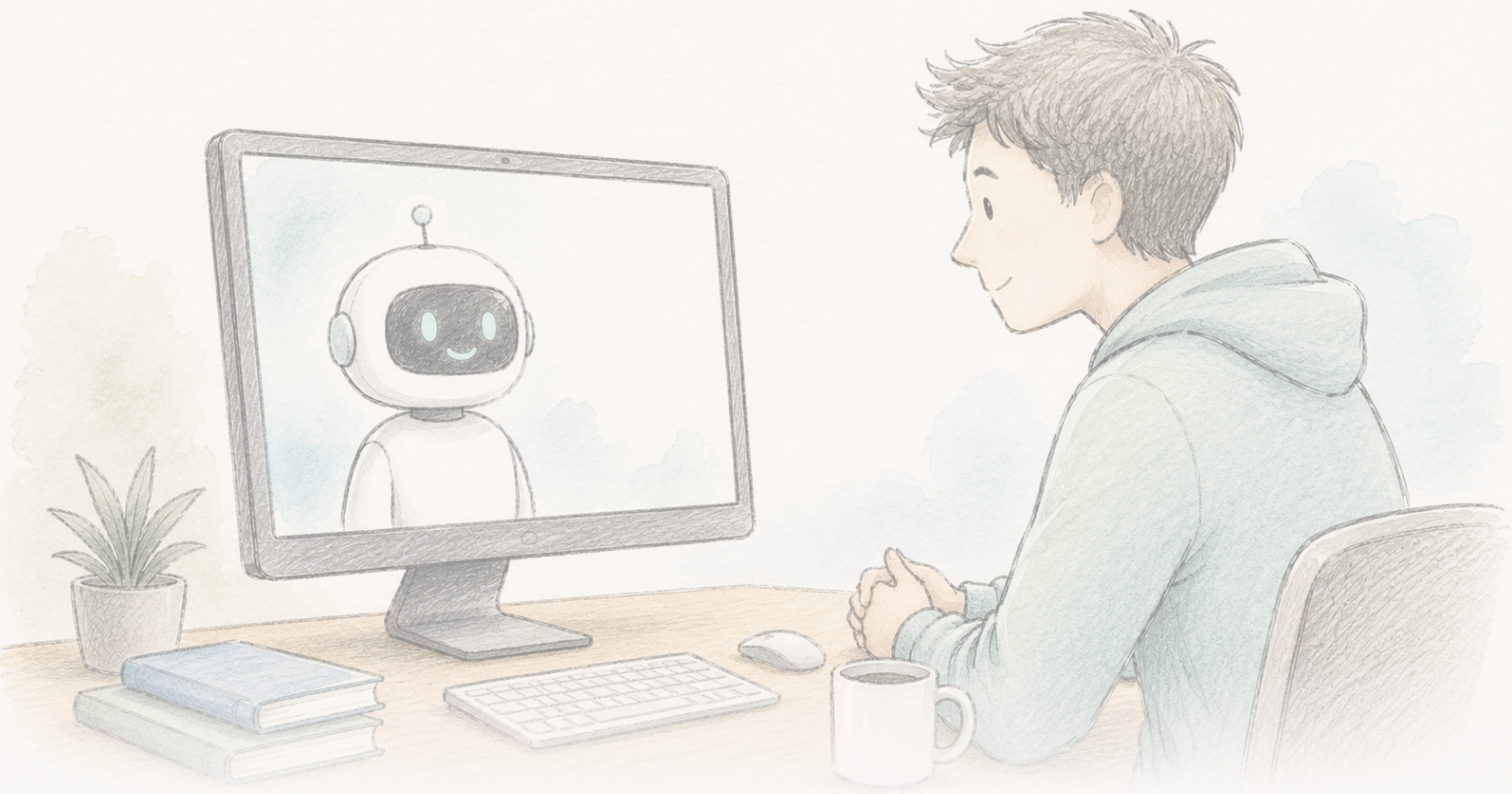




The CLEAN PAPER

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

챗봇 대화가 음모론 믿음을 몇 달 동안 줄인 것으로 나타났다

2024 년 연구는 GPT-4 Turbo 와 세 차례 대화한 뒤 참가자가 믿던 음모론에 대한 믿음이 약 20% 줄었고, 효과가 두 달 지속되며 여러 음모론에 일반화됐고 AI 의 사실 주장도 압도적으로 정확했다고 보고했다. 2026 년 6 월 데이터 처리와 재현성 문제 때문에 논문에는 *Editorial Expression of Concern* 이 붙었다. 저자들은 수정 분석에서도 결과의 방향·유의성·크기가 유지된다고 말하며 *Science* 는 아직 평가 중이다. 실제로 흥미롭고 신중하게 설계된 결과지만 현재 검토 중이며, 확정도 반박도 아니다.

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 는 데이터 처리와 재현성 문제를 평가하는 동안 이 논문에 Editorial Expression of Concern 을 붙였다. 이는 철회도 연구부정 판정도 아니다. 검토가 해결될 때까지 보고된 결과를 잠정적으로 읽어야 한다는 뜻이다.

Editorial Expression of Concern →

결국은 사실 — 그리고 논문에 붙은 경고 표시

2024 년 한 연구는 편만한 통념과 어긋나는 결과를 보고했다. 자신이 실제로 믿는 음모론과 그 근거를 직접 제시한 사람에게, 그 구체적인 증거에 답하도록 프롬프트된 GPT-4 Turbo 와 세 차례 짧게 대화하게 하면 평균적으로 그 음모론에 대한 믿음이 약 20% 줄었다. 감소는 두 달 뒤에도 남아 있었다. 존 F. 케네디 암 살, 외계인, 일루미나티 같은 고전적 음모론부터 코로나 19 와 2020 년 미국 대선 같은 시사 음모론까지 폭 넓게 나타났고, 믿음이 강하거나 정체성과 연결된 참가자에게서도 효과가 있었다. 전문 팩트체커가 AI 가 한 사실 주장 128 개를 평가했을 때 127 개는 참, 1 개는 오해를 부를 수 있는 표현, 거짓은 0 개였다.

널리 퍼진 헤드라인은 '토끼굴에서 사람을 말로 끌어낼 수 있다' 는 것이었다. 음모론 신봉자도 증거의 영향을 받을 수 있으며, 단지 그 사람이 실제로 드는 근거에 충분히 구체적으로 답해 주는 증거가 필요하다는 이야기다. 실제로 흥미로운 주장이고, 연구 설계도 설득 연구치고는 상당히 신중했다.

그런데 2026 년 6 월 *Science* 는 이 논문에 **Editorial Expression of Concern(편집부 우려 표명)** 을 붙였다. 대부분의 재전달에서 빠질 부분이지만, 지금 이 결과가 얼마만큼의 무게를 견딜 수 있는지를 결정하는 바로 그 부분이다.

Expression of Concern 은 무엇이며 무엇이 아닌가

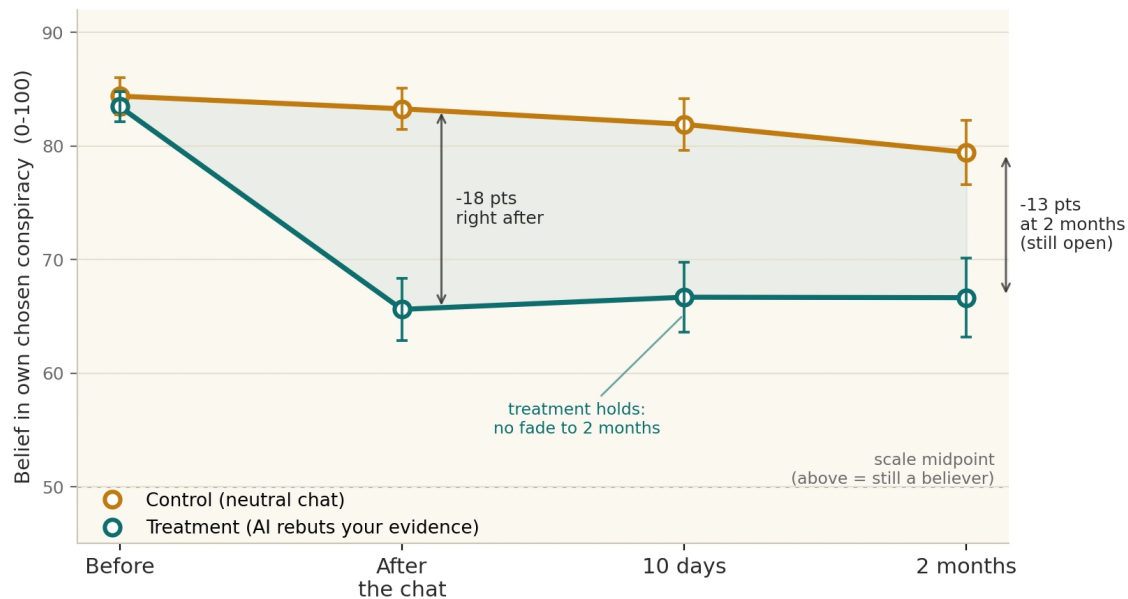
Editorial Expression of Concern 은 논문에 문제가 제기됐지만 아직 조사·평가가 끝나지 않았을 때 저널이 독자에게 알리기 위해 붙이는 공식 경고 표시다. **철회 (retraction) 가 아니다**. 논문이 취소된 것도 아니고, 그 자체가 연구부정행위 판정도 아니다. 의미는 '기록이 바뀔 수 있으니 이 주의 사항과 함께 읽으라' 는 것이다. 이번 경우 저자들은 문제가 알려진 뒤 자체 조사를 했고 구체적인 내용을 *Science* 에 보고했으며 수정된 분석도 제출했다.

지적된 문제는 구체적이다. 참가자 선별 기준이 원고와 공개 분석 코드 사이에서 일관되게 적용되지 않았고, 공개 데이터셋에는 코드 병합 오류로 불필요한 행들이 끼어 있었다. 이 때문에 공개 자료만으로 일부 보고 수치를 재현하기 어려웠다. 저자들은 수정된 분석 파이프라인과 업데이트된 결과를 *Science* 에 제출했고, 방향·통계적 유의성·실질적 크기에서 원래 결과와 일치한다고 말한다. 현재 *Science* 가 이를 평가 중이다. 평가가 끝날 때까지 정확한 수치에는 저널만이 없앨 수 있는 물음표가 붙어 있다.

따라서 이 이야기는 동시에 두 가지다. 사실이 굳어진 믿음을 움직일 수 있는가에 대한 흥미로운 결과이고, 과학 기록이 공개적으로 자기 자신을 교정하는 실제 사례다. 이 글의 목적은 둘을 섞지 않는 데 있다.

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.

연구에서는 참가자가 직접 제시한 근거를 반박하는 세 차례 AI 대화가 그 사람이 선택한 음모론에 대한 믿음을 평균 약 20% 줄였다고 보고했다. 관계없는 주제로 대화한 대조군과 달리 감소는 10 일과 2 개월 뒤에도 측정됐다. 효과는 지속적이었지만 부분적이었다. 대부분 참가자는 이후에도 음모론을 믿었고, 믿음이 줄어든 것이지 사라진 것은 아니다. 현재 이 논문에는 보고 불일치와 공개 데이터셋 오류 때문에 *Science*의 Editorial Expression of Concern(2026년 6월)이 붙어 있다. 저자들은 수정 분석에서도 효과의 방향·유의성·크기가 유지된다고 말하지만 *Science*의 평가는 진행 중이다. 이 차트는 The Clean Paper가 해당 공개 데이터셋에서 재구성했으므로 표시 값은 잠정적이며 논문의 최종 수치가 아니다.

Original figure —The Clean Paper CC BY 4.0

연구진이 한 일

- 참가자 **2,190 명**을 대상으로 두 실험을 진행했다. 각 참가자는 자신이 믿는 음모론과 그것을 뒷받침한다고 생각하는 증거를 자기 말로 적었다.
- 각 참가자는 GPT-4 Turbo 와 세 차례 대화했다. **처치 조건**에서는 AI가 참가자가 제시한 구체적 근거를 반박하도록 프롬프트됐고, **대조 조건**에서는 관계없는 주제를 논의했다.
- 선택한 음모론에 대한 믿음을 대화 전과 직후에 측정하고, **10 일 뒤와 2 개월 뒤** 다시 연락해 추적했다.
- 전문 팩트체커가 표본에서 AI가 한 사실 주장 128 개 모두의 정확성을 평가했다.
- 효과가 다른 관계없는 음모론으로도 번지는지, 그리고 특이성 검사를 위해 실제로 **사실인 음모적 사건**에 대한 믿음까지 떨어뜨리는지도 확인했다.

무엇을 발견했나

- 처치군에서 대조군 대비 선택한 음모론에 대한 믿음이 **평균 약 20% 감소**했다 (연구 1: 100 점 척도에서 95% 신뢰구간 [13.8, 19.7] 점, $P < 0.001$).
- **효과는 지속됐다.** 처치군의 믿음 감소는 다시 올라오지 않았고 10 일과 2 개월 뒤에도 대화 직후 수준 근처에 머물렀다. 대조군은 계속 더 높았다. 논문은 선택한 음모론에 대한 효과가 최소 2 개월 동안 줄어들지 않고 지속됐다고 기술한다.
- **여러 유형의 음모론에 일반화됐고,** 처음 믿음이 강한 참가자에게서도 나타났다.
- 이 환경에서 **AI 의 사실 주장은 매우 정확했다.** 128 개 중 127 개 (99.2%) 가 참, 1 개 (0.8%) 가 오해를 부를 수 있는 표현, 거짓은 없었다.
- 단순한 전반적 의심 증가가 아니라 **특이적 효과**였다. 실제로 사실인 음모적 사건에 대한 믿음은 줄지 않았다. 즉 모든 것을 냉소적으로 만들기보다 근거가 약한 믿음을 움직였음을 시사한다.
- **작은 파급효과**도 있었다. 다른 관계없는 음모론에 대한 믿음은 약 12% 줄었고, 참가자들은 음모론 주장에 반박할 의도가 더 커졌다고 보고했다. 핵심 효과는 연구 2 에서도 유지됐고, 대조군이 없는 더 작은 표본에서도 같은 방향을 보였다 (더 약한 증거지만 일관됨).

이 연구가 증명하지 않는 것

- 지금 **아무 의문 없이 확정된 결과는 아니다.** 데이터 처리와 재현성 문제로 Editorial Expression of Concern 이 붙어 있고, 수정 수치는 아직 *Science* 가 평가 중이다. 평가가 끝날 때까지 구체적인 값은 잠정적으로 읽어야 한다.
- AI 가 음모론 믿음을 '치료' 한다는 것을 **보여주지 않는다.** 평균 약 20% 감소는 의미 있는 변화지만 제거는 아니다. 대부분은 이후에도 음모론을 믿되 덜 믿었다.
- **실제 환경 배치 결과가 아니다.** 참가자들은 연구에 자발적으로 참여해 AI 와 성실히 대화했다. 현실에서는 음모론에 가장 강하게 헌신한 사람이 자신과 논쟁할 챗봇을 일부러 찾을 가능성이 가장 낮을 수 있다.
- 99.2% 정확도는 **이 과업, 이 모델, 이 신중한 프롬프트에 한정**된다. 언어 모델이 일반적으로 참말을 한다는 보장이 아니다. 저자들은 같은 개인화 설득 능력이 거짓 믿음을 밀도록 사용되면 반대 방향으로도 작동할 수 있다고 명시한다.
- 여기서 '지속적' 은 **두 달**을 뜻한다. 한 번의 대화치고 인상적이지만 영구적이라는 뜻은 아니다.

근거는 얼마나 탄탄한가?

- **연구 설계는 실제 강점이다.** 처치-대조 통제실험, 위약에 가까운 명확한 대조 조건, 두 연구와 독립 표본에서의 반복, 지속성 추적, AI 주장 자체의 정확성 검사까지 포함했다. 설득 연구 대부분보다 신중했고 시험된 곳에서는 효과 방향이 일관됐다.
- **하지만 Expression of Concern 은 각주가 아니다.** 공개된 데이터와 코드에서 보고 수치를 재생성할 수 있는가는 결과의 신뢰성 일부이고, 바로 그 부분에 문제가 생겼다. 선별 기준 불일치와 코드 병합 오류로 중복·불필요한 행이 들어갔다. 수정 파이프라인에서도 결과가 유지된다는 저자들의 설명은 고무적이고 그럴듯하지만 아직 저자들의 설명일 뿐 독립 판정이 아니다. 기다려야 할 것은 *Science* 의 평가

다.

- **정직한 현재 상태는 '문제 표시됨 (flagged)' 이다. '반박됨' 도 '확인됨' 도 아니다.** 설계가 좋고 눈에 띄는 연구지만 정확한 숫자가 공식 검토 중이다. 좋은 이야기를 이런 불편한 지점에 남겨두는 것이 정확한 태도다.

왜 중요한가

수년 동안 영향력 있는 견해 하나는 음모론 신봉자의 믿음이 심리적 필요와 정체성에 의해 움직이므로 사실로 논박해도 바뀌지 않는다는 것이었다. 이 결과가 유지된다면, 지속적인 증거 기반 감소는 그런 비관론에 중요한 균형추가 된다. 일반적인 팩트체크가 신봉자가 실제로 드는 구체적인 증거에 대응하지 못했던 것이 문제의 일부였고, LLM 은 이를 대규모로 개인화할 수 있다는 가능성을 보여준다.

과학이 어떻게 작동해야 하는지 보여주는 사례로도 중요하다. Expression of Concern 의 교훈은 유명한 결과가 무가치하다는 것이 아니다. 오류가 드러나고, 데이터가 다시 검토되고, 주장이 공개적으로 재확인되는 과정이 작동한다는 것이다. 그 장치가 움직이는 것은 스캔들이 아니라 기능이다. 그래서 지금 필요한 태도는 박수도 무시도 아니라 인내다.

AI 측면에서도 양날이다. 맞춤형 논증으로 거짓 믿음을 약화시키는 같은 능력은 거짓 믿음을 강화하는 데도 사용할 수 있다. 기법은 중립적이지만 목적은 그렇지 않다.

핵심 요약

2024 년 연구는 GPT-4 Turbo 와 세 차례 대화한 뒤 참가자가 믿던 음모론에 대한 믿음이 약 20% 줄었고, 효과가 2 개월 지속되며 여러 음모론에 나타났고 AI 의 사실 주장도 압도적으로 정확하게 평가됐다고 보고했다. 그러나 2026 년 6 월 데이터 처리와 재현성 문제 때문에 이 논문에 Editorial Expression of Concern 이 붙었다. 저자들은 수정 분석에서도 효과의 방향·유의성·크기가 유지된다고 보고했고, *Science* 는 아직 평가 중이다. 따라서 현재 가장 정확한 해석은 '정말 흥미롭고 신중하게 설계된 결과가 공식 검토 중' 이라는 것이다. 확정도 반박도 아니다. 이 연구에 가장 정직한 문장은 과학 과정 자체가 쓰고 있다. **다시 확인하라.**

출처

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>