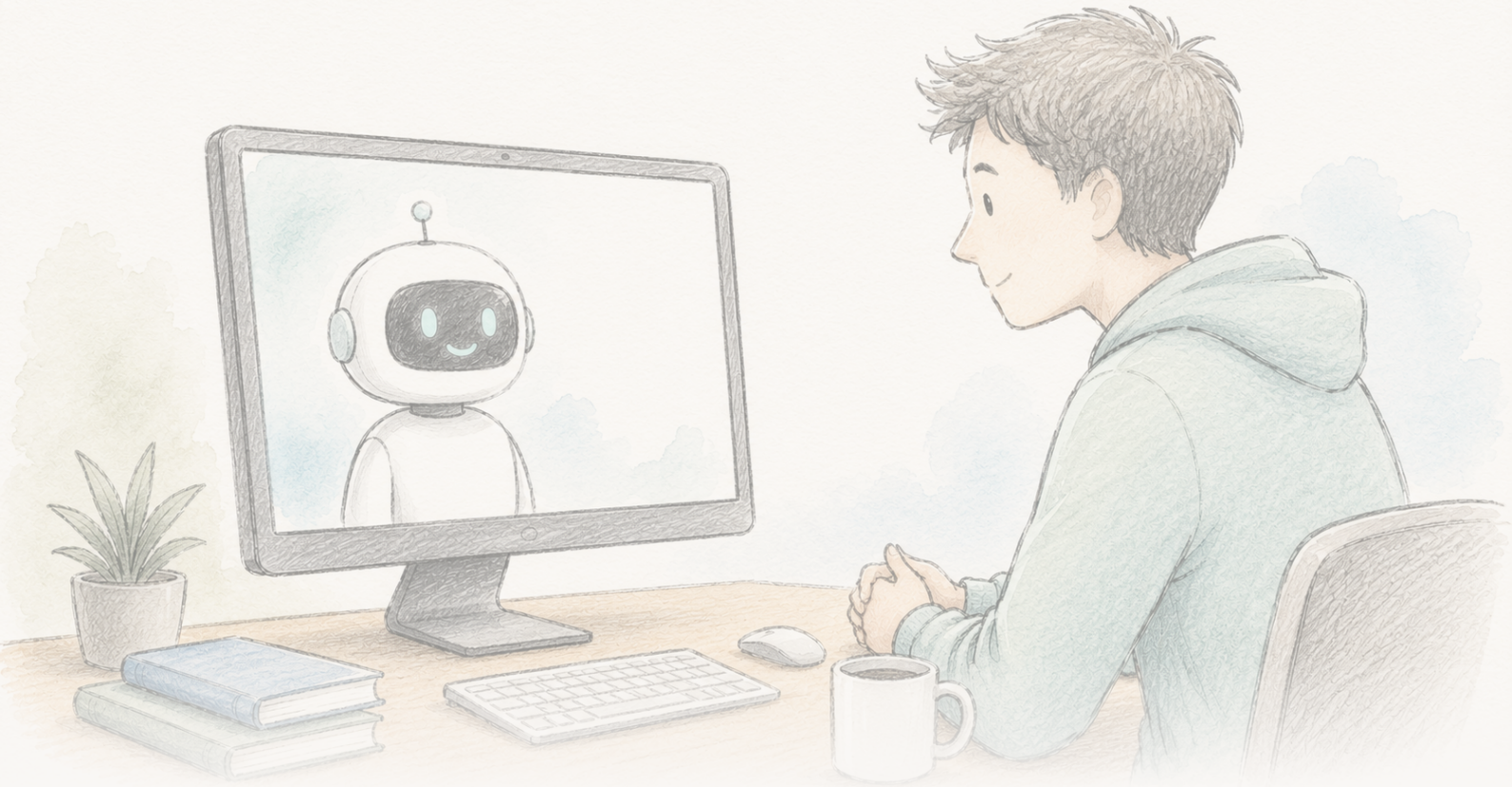


The Clean Paper

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

एक चैटबॉट ने षड्यंत्र-विश्वास को महीनों तक घटाया—कम-से-कम प्रारंभिक नतीजों में

2024 के एक अध्ययन में GPT-4 Turbo के साथ तीन दौर की बातचीत के बाद लोगों के अपने चुने हुए षड्यंत्र-सिद्धांत पर विश्वास में औसतन लगभग 20% कमी आई। प्रभाव दो महीने तक बना रहा, अलग-अलग षड्यंत्रों में दिखा और AI के तथ्यात्मक दावों को लगभग पूरी तरह सही आँका गया। जून 2026 में डेटा-प्रबंधन और पुनरुत्पादन से जुड़ी समस्याओं के कारण शोधपत्र पर Editorial Expression of Concern लगाया गया। लेखक कहते हैं कि सुधारे गए विश्लेषण में प्रभाव की दिशा, सांख्यिकीय महत्त्व और आकार बने रहते हैं; Science अभी इसकी जाँच कर रहा है। परिणाम सचमुच दिलचस्प और सावधानी से बनाया गया है — लेकिन फिलहाल समीक्षा के अधीन है, न पूरी तरह स्थापित और न खारिज।

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 ने डेटा-प्रबंधन और पुनरुत्पादन के प्रश्नों की जाँच के दौरान इस शोधपत्र पर Editorial Expression of Concern लगाया है। यह न वापसी (retraction) है, न कदाचार का निष्कर्ष। इसका अर्थ इतना है कि समीक्षा पूरी होने तक रिपोर्ट किए गए परिणामों को अस्थायी माना जाना चाहिए।

Editorial Expression of Concern →

आखिरकार तथ्य — और शोधपत्र पर एक औपचारिक चेतावनी

2024 में एक अध्ययन ने ऐसा परिणाम रिपोर्ट किया जो एक सुविधाजनक पारंपरिक wisdom के खिलाफ जाता है। किसी व्यक्ति को AI — GPT-4 Turbo — के साथ तीन rounds की छोटी बातचीत करवाई, जहाँ AI उस षड्यंत्र सिद्धांत के समर्थन में व्यक्ति द्वारा खुद दिए गए विशिष्ट साक्ष्य का जवाब देता है, और औसतन उस सिद्धांत पर उसका विश्वास लगभग 20% घटता है। दो महीने बाद भी गिरावट बनी रही। प्रभाव classic conspiracies (John F. Kennedy की हत्या, aliens, Illuminati) से लेकर topical ones (COVID-19, 2020 US election) तक दिखा, और उन लोगों में भी जिनका विश्वास मजबूत और identity से जुड़ा था। एक professional fact-checker ने AI के 128 factual दावों को grade किए: 127 सत्य, एक misleading, कोई असत्य नहीं।

फैलने वाला सुरंगी था कि लोगों को rabbit hole से बाहर *बात करके* निकाला जा सकता है — षड्यंत्र believers साक्ष्य की पहुँच से बाहर नहीं, उन्हें बस सही साक्ष्य पर्याप्त specificity के साथ चाहिए। यह सचमुच दिलचस्प दावा है, और अध्ययन persuasion शोध के बहुत से कम से अधिक सावधानी से बनाई गई थी।

फिर जून 2026 में *Science* ने शोधपत्र पर **Editorial अभिव्यक्ति of Concern** लगा दी। अधिकांश retellings इस हिस्से को छोड़ देंगी, जबकि अभी परिणाम पर कितना weight रखा जा सकता है, यह तय करने वाला हिस्सा यही है।

Expression of Concern क्या है — और क्या नहीं

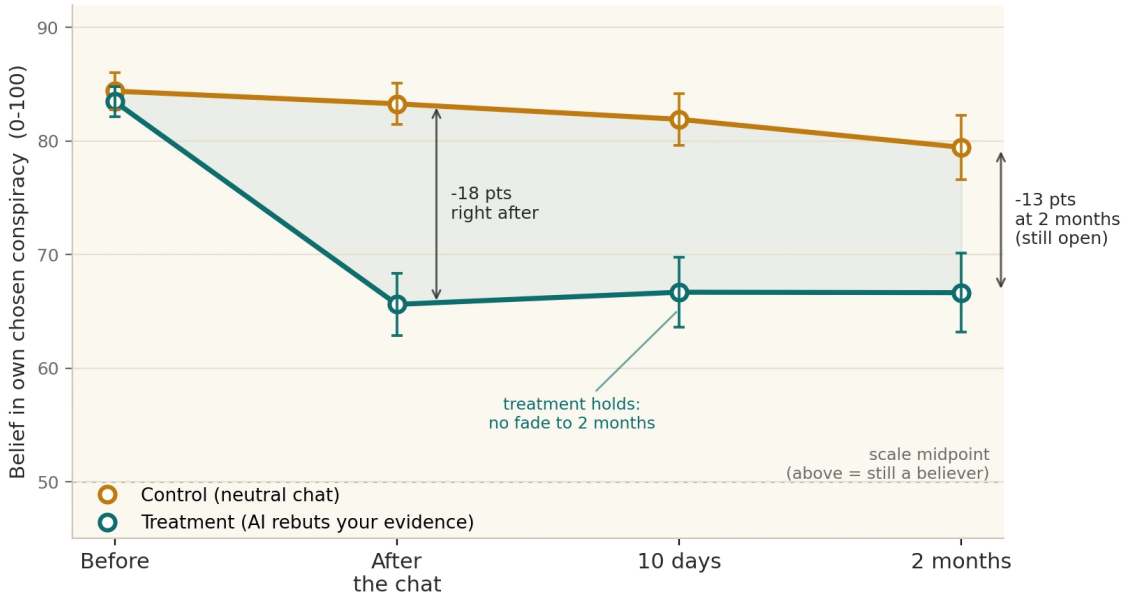
Editorial अभिव्यक्ति of Concern वह औपचारिक संकेतक है जिसे journal किसी शोधपत्र पर तब लगाता है जब उसके बारे में सवाल उठे हों और जाँच अभी जारी हो। यह **retraction नहीं** है — शोधपत्र वापस नहीं लिया गया — और अपने-आप में misconduct का निष्कर्ष भी नहीं। इसका अर्थ है: इस caveat को ध्यान में रखकर पढ़ें, क्योंकि scientific रिकॉर्ड बदल सकता है। यहाँ issues की जानकारी मिलने के बाद लेखकों ने जाँच की विशिष्ट बातें *Science* को रिपोर्ट किए और corrected विश्लेषण दिया।

संकेतक किया गया issue विशिष्ट है। लेखक को participant-जाँच criteria manuscript और प्रकाशित विश्लेषण कोड में अलग तरह से apply होने की inconsistencies के बारे में बताया गया, और साझा जनकि डेटासेट में code-merging त्रुटि से अतिरिक्त rows जुड़ गई थीं — ऐसी समस्याएँ जिनके कारण released पदार्थ से कुछ रिपोर्ट किया संख्याएँ पुनरुत्पादित करना कठिन था। लेखकों ने *Science* को corrected विश्लेषण pipeline और updated परिणाम दिए हैं, जिनके बारे में उनका कहना है कि वे original से **दिशा, सांख्यिकीय महत्त्व और substantive आकार** में मेल खाते हैं। *Science* उनका मूल्यांकन कर रहा है। जब तक वह प्रक्रिया पूरा नहीं होती, सटीक चित्र पर प्रश्न mark है जिसे केवल journal हटा सकता है।

इसलिए यहाँ एक साथ दो कहानियाँ हैं: entrenched beliefs को तथ्य बदल सकते हैं या नहीं, इस पर intriguing परिणाम; और scientific रिकॉर्ड का खुद को खुले में सही करने का जीवंत उदाहरण। इस लेख का उद्देश्य दोनों को अलग रखना है।

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.

अध्ययन में प्रतिभागीद्वाराखुद बताए सबूत कोrebut करने वालीतीन-दौर AI बातचीत ने चुनीषड्यंत्र पर मान्यताऔसतन लगभग 20% घटाई; unrelated-topic नियंत्रण chat के विपरीत यह गिरावट 10 days और 2 months बाद भीमापने योग्य थी। रिपोर्ट किया प्रभाव durable लेकिन partial था अधिकांश प्रतिभागीबाद में भीषड्यंत्र मानते थे — कमी cure नहीं। शोधपत्र पर फिलहाल reporting inconsistencies और सार्वजनिक डेटासेट कीत्रुटि के कारण Editorial अभिव्यक्ति of Concern (*Science*, June 2026) है; लेखक कहते हैं corrected विश्लेषण प्रभाव कीदिशा महत्त्व और आकार बनाए रखताहै, जबकि *Science* अभीevaluate कर रहा है। यह चार्ट साफ़ शोधपत्र ने उसीसार्वजनिक डेटासेट से पुनर्निर्मित कियाहै, इसलिए दिखाए गए अस्थायीहैं, शोधपत्र के अंतिम चित्र नहीं।

Original figure — The Clean Paper CC BY 4.0

शोधकर्ताओं ने क्याकिया

- 2190 प्रतिभागीके साथ दोप्रयोग चलाए। हर प्रतिभागीने अपने शब्दोंमें ऐसीषड्यंत्र सिद्धांत बताई जिस पर वह विश्वास करताथा और वह साक्ष्य भीबताया जोउसके हिसाब से उसे समर्थन करताथा।
- हर प्रतिभागीने GPT-4 Turbo के साथ तीन-round बातचीत की। **उपचार** स्थिति में AI कोप्रतिभागीके विशिष्ट साक्ष्य कारebuttal देने के लिए प्रॉम्प्ट कियागया; **नियंत्रण** स्थिति में उसने unrelated topic discuss किया।
- चुनीषड्यंत्र पर मान्यताबातचीत से पहले और तुरंत बाद मापा; फिर प्रतिभागीको **10 days और 2 months** पर दोबारा संपर्क किया।
- Professional fact-checker से नमूनामें AI द्वाराकिए गए सभी128 factual दावे कीसटीकताevaluate कराई।
- जाँचाकि प्रभाव दूसरी unrelated conspiracies तक spill over होताहै यानहीं— और specificity परीक्षण के रूप में यह भीकि कहींintervention उन conspiracies पर मान्यतातोनहींघटताजोवास्तव में **सत्य** हैं।

क्यामलिा

- उपचार समूह में नियंत्रण कीतुलनामें चुनीषड्यंत्र पर मान्यतामें **लगभग 20% औसत कमी**(अध्ययन 1: 100-बिंदु पैमानापर 95% विश्वास-अंतराल [13.8, 19.7] बिंदु, $P < 0.001$)।
- **प्रभाव कायम रहा।** उपचार समूह में गिरावट fade नहींहुई: मान्यता10 days और दोmonths पर भीchat के तुरंत बाद के level के पास रही, जबकि नियंत्रण अधिक रहा। शोधपत्र इसे चुनीषड्यंत्र पर कम-से-कम दोmonths तक undiminished प्रभाव बताताहै, momentary wobble नहीं।
- **प्रभाव अलग-अलग षड्यंत्र प्रकारोंतक भीफैला,** और उन प्रतिभागीमें भीरहाजिनकाप्रारंभिक मान्यतामजबूत था।
- **इस परिस्थिति में AI के दावे accurate थे:** 128 में 127 (99.2%) सत्य, 1 (0.8%) misleading, none असत्य।
- **यह विशिष्ट था,** blanket doubt नहीं intervention ने सत्य conspiracies पर मान्यताकम नहींकिया, जिससे संकेत मिलताहै कि unsupported beliefs कदम हुई, लोग हर चीज़ के प्रति cynical नहींबने।
- **मध्यम spillover** भीहुआ — other unrelated conspiracies पर मान्यतालगभग 12% घटाऔर प्रतिभागीने षड्यंत्र दावे काpushback करने कीअधिक intention रिपोर्ट की। मुख्य प्रभाव अध्ययन 2 में भीरहाऔर बिनानियंत्रण समूह वाले छोटे नमूनामें भीसमान दिशामें गया(और कमज़ोर संकेत, लेकिन संगत)।

यह क्यासंभव नहींकरता

- अभीयह **बिनासवाल वालापरिणाम नहीं**है। शोधपत्र data-handling और reproducibility समस्याएँ के कारण Editorial अभिव्यक्ति of Concern के तहत है, और corrected संख्याएँ कामूल्यांकन *Science* कर रहाहै। समीक्षापूराहोने तक विशिष्ट मान कोअस्थायीमानें।
- यह **AI षड्यंत्र मान्यता"cure" करताहै** नहींदिखाता। ~20% औसत कमीसार्थक बदलाव है, erasure नहीं— अधिकांश प्रतिभागीबाद में भीसंदिग्ध मानते थे, बस कम।
- यह **वास्तविक दुनियातैनातीपरिणाम नहीं**है। प्रतिभागीअध्ययन में opt-in करके AI के साथ अच्छाfaith में engaged हुए; वास्तविक दुनियामें षड्यंत्र के सबसे committed believers शायद उस chatbot कोखोजने हीकम जाएँ जोउनसे तर्क देनाकरे।
- 99.2% सटीकता**इस कार्य, मॉडल और संकेत prompting तक सीमित** है — भाषामॉडल के हमेशासत्य होने कीसमान्य गारंटीनहीं। लेखक स्पष्ट कहते हैं कि यहीpersonalized persuasive शक्ति अगर असत्य beliefs कीओर लक्ष्य की जाए तोउन्हें भीबढ़ासकतीहै।
- "Durable" यहाँ**दोmonths** है, जोएकल बातचीत के लिए impressive है, permanent के बराबर नहीं।

प्रमाण कितनामजबूत है

- **डिज़ाइन वास्तविक strength है।** नियंत्रित treatment-versus-नियंत्रण प्रयोग, साफ़ प्लेसबोstyle नियंत्रण, दो अध्ययन और स्वतंत्र नमूनामें प्रतिकृति निर्माण, durability follow-ups, और AI के own दावे पर स्पष्ट सटीकता जाँच — persuasion शोध के बहुत से कम से यह अधिक संकेत है, और जहाँजहाँपरीक्षण हुआ प्रभाव कीदिशासंगत रही।
- **लेकिन अभिव्यक्ति of Concern footnote नहीं है।** Released डेटाऔर कोड से रिपोर्ट कियासंख्याएँ regenerate कर पानापरिणाम कीtrustworthiness काहिस्साहै, और यहीचीज़ टूटी— जाँच-criteria inconsistencies और code-merging त्रुटि से duplicated/अतिरिक्त rows। लेखकोंकाकहनाकि corrected pipeline परिणाम बचतीहै encouraging और संभावित है, लेकिन फिलहाल वह लेखकोंकाown account है, स्वतंत्र verdict नहीं। *Science* कामूल्यांकन निर्णायक

अगलाचरण है।

- **ईमानदार स्थिति "flagged" है, "debunked" या "पुष्ट" नहीं।** एक well-built, striking अध्ययन जिसके सटीक संख्याएँ औपचारिक समीक्षा में हैं। अच्छी कहानी को यहीं छोड़ना असहज है, लेकिन फिलहाल accurate स्थिति यही है।

यह क्यों महत्वपूर्ण है

सालों तक एक प्रभावशाली view था कि षड्यंत्र believers psychological needs और identity से driven होते हैं, इसलिए तथ्य से उन्हें मान्यता से बहर तर्क देना नहीं किया जा सकता। अगर durable साक्ष्य-based कमी टिकती है, तो वह उस pessimism के खिलाफ सार्थक counterweight है: शायद समस्या का एक हिस्सा यह था कि generic debunking कभी उस *विशिष्ट* साक्ष्य से engage नहीं करती जिसे believer वास्तव में cite करता है — और LLM ऐसा पैमाने पर कर सकता है।

यह इस बात का मामला अध्ययन भी है कि science को काम कैसे करना चाहिए। अभिव्यक्ति of Concern का सबक यह नहीं कि celebrated परिणाम बेकार हैं; सबक यह है कि त्रुटियाँ सतह होती हैं, डेटा फिर examine होता है, और दावे सार्वजनिक में re-check होते हैं। यह मशीनरी चलना विशेषता है, scandal नहीं— और इसी वजह से परिणाम पर सही posture applause या dismissal नहीं, patience है।

AI के बारे में भी बात दो श्रोताओं में जाती है। Tailored तर्क से असत्य मान्यता कम करने की वही क्षमता किसी असत्य मान्यता को उतनी ही व्यावहारिक रूप से बढ़ा सकती है। Technique तटस्थ है; aim तटस्थ नहीं।

संक्षिप्त सार

2024 के एक अध्ययन ने पाया कि GPT-4 Turbo के साथ तीन-दौर बातचीत ने लोगों की मानी हुई षड्यंत्र सिद्धांत पर मान्यता लगभग 20% घटाई, प्रभाव दो months तक रहा और अलग-अलग षड्यंत्र प्रकारों में भी बना रहा। AI के factual दावे overwhelming रूप से accurate grade किए गए। जून 2026 में data-handling और reproducibility समस्याएँ के कारण शोधपत्र पर Editorial अभिव्यक्ति of Concern लगाई गई; लेखक रिपोर्ट करते हैं कि corrected विश्लेषण निष्कर्ष की दिशा, महत्त्व और आकार बनाए रखता है, और *Science* अभी evaluate कर रहा है। इसे वास्तव में दिलचस्प, carefully built परिणाम की तरह पढ़ें जो फिलहाल औपचारिक समीक्षा में है — तय नहीं, debunked नहीं।

स्रोत

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`