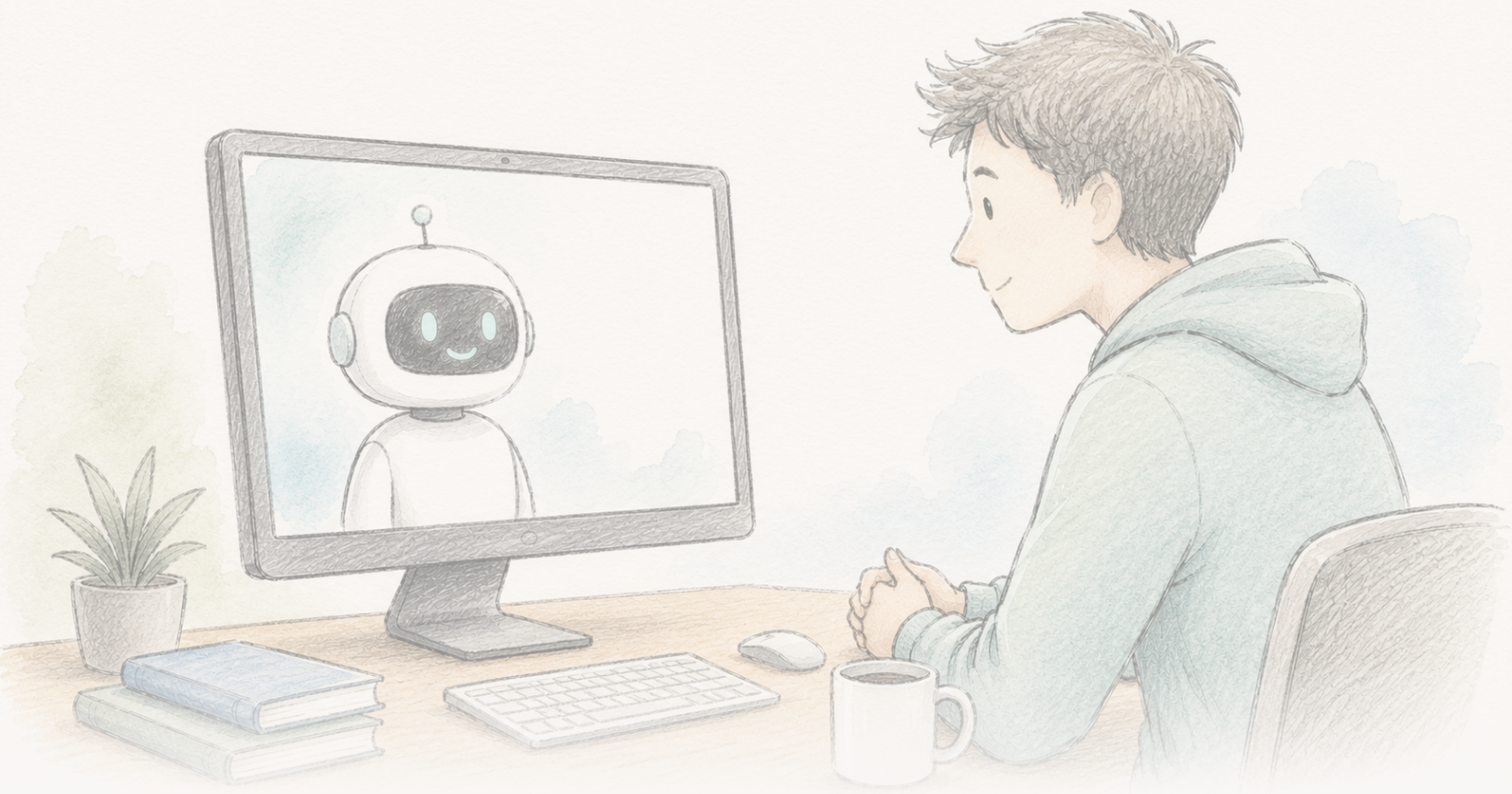




The CLEAN PAPER

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

एक चैटबॉट ने षड्यंत्र-वश्विस्स कोमहीमोतक घटया— कम-से-कम प्रारंभकि नतीजोमें

2024 के एक अध्ययन में GPT-4 Turbo के साथ तमि दौ कीबसचित के बढ लगेके अपने चुने हुए षड्यंत्र-सदिधंस पर वश्विस्स में औसतन लगभग 20% कमीआई। प्रभस दोमहीमे तक बनारहा अलग-अलग षड्यंत्रोमें दखा और AI के तथ्यसुमक दमोकोलगभग पूरीतरह सहीआँकागया। जून 2026 में डेटाप्रबंधन और पुनरुत्पदन से जुडीसमस्योके कारण शेषपत्र पर Editorial Expression of Concern लगयागया। लेखक कहते हैं कि सुधरे गए वश्विलेषण में प्रभस कीदशा संख्यकीय महत्त्व और आकर बने रहते हैं; Science अभीइसकीजाँच कर रहाहै। परणिस सचमुच दलिचस्प और सभधमीसे बनयागयाहै — लेकनि फलिहल समीक्षाके अधिम है, न पूरी तरह स्थमति और न खरजि।

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 डेटाप्रबंधन और पुनरुत्पन्न के प्रश्नों की जाँच के दौरान इस शोधपत्र पर Editorial Expression of Concern लगाया है। यह न वसति (retraction) है, न कदाचित् कानिष्कर्ष। इसका अर्थ इतना है कि समीक्षा पूरी होने तक रिपोर्ट किए गए परिणमों को अस्थायी माना जा रहा है।

Editorial Expression of Concern →

आखिरकार तथ्य — और शोधपत्र पर एक औपचारिक चेतनी

2024 में एक अध्ययन ने ऐसा परिणम रिपोर्ट किया जो एक सुविधानक पारंपरिक wisdom के खिलाफ जाता है। किसी व्यक्ति को AI — GPT-4 Turbo — के साथ तीन rounds की छोट्टी बातचीत करवाए, जहाँ AI उस षड्यंत्र सिद्धि के समर्थन में व्यक्ति द्वारा खुद दिए गए विशिष्ट सक्षय का जवाब देता है, और औसतन उस सिद्धि पर उसका विश्वास लगभग 20% घटता है। दोमहीमें बढ भी गिरा नही बन रही। प्रभम classic conspiracies (John F. Kennedy की हत्या, Illuminati) से लेकर topical ones (COVID-19, 2020 US election) तक दिखा और उन लोगों में भी जनिका विश्वास मजबूत और identity से जुड़ा था। एक professional fact-checker ने AI के 128 factual दमों को grade किया: 127 सत्य, एक misleading, कोई असत्य नहीं।

फैलने वाला सुखी थाक लोगों को rabbit hole से बहर बत करके निकाला जा सकता है — षड्यंत्र believers सक्षय की पहुँच से बहर नहीं उन्हें बस सही सक्षय पर्याप्त specificity के साथ चाहिए। यह सचमुच दिलचस्प दम है, और अध्ययन persuasion शोध के बहुत से कम से अधिक सम्मानी से बनई गई थी।

फरि जून 2026 में Science ने शोधपत्र पर Editorial अभिव्यक्ति of Concern लगा दी। अधिकांश retellings इस हिससे को छोड़ देंगी जबकि अभी परिणम पर कतिना weight रखा जा सकता है, यह तय करने वाला हिसाब ही है।

Expression of Concern क्या है — और क्या नहीं

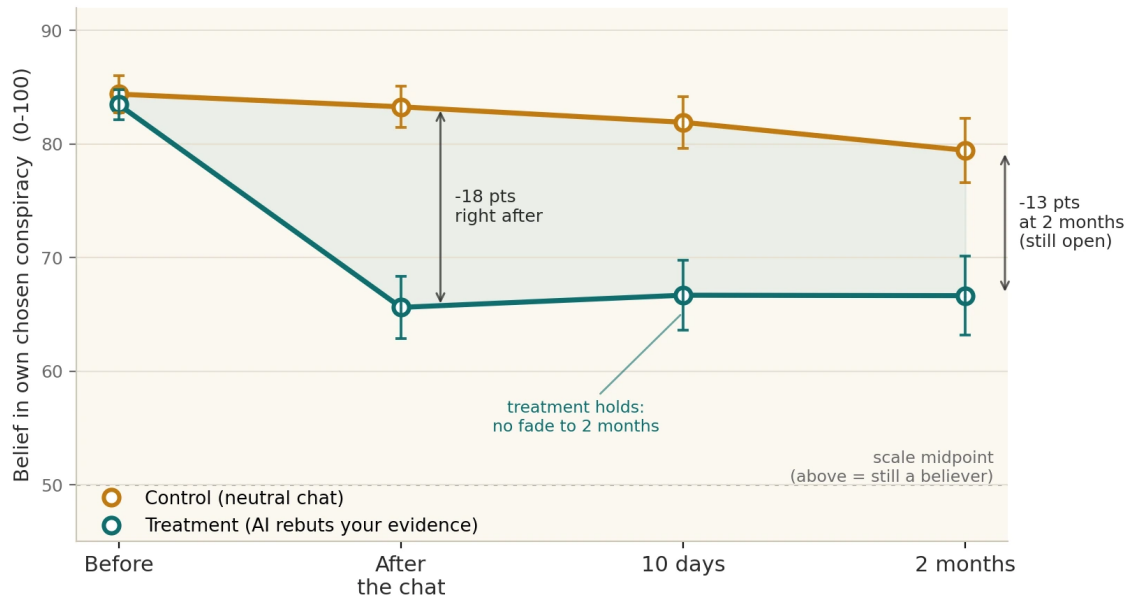
Editorial अभिव्यक्ति of Concern वह औपचारिक संकेतक है जिसे journal किसी शोधपत्र पर तब लगाता है जब उसके बारे में सवाल उठे हो और जाँच अभी जारी हो। यह retraction नहीं है — शोधपत्र वापस नहीं लिया गया — और अपने-आप में misconduct कानिष्कर्ष भी नहीं। इसका अर्थ है: इस caveat को ध्यान में रखकर पढ़ें, क्योंकि scientific रिकॉर्ड बदल सकता है। यहाँ issues की जमकरी मिलने के बाद लेख कोने 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 सर्वजनिक डेटासेट में code-merging त्रुटि से अतिरिक्त rows जुड़ गई थी — ऐसी समस्याएँ जिनके कारण released पदार्थ से कुछ रिपोर्ट किया संख्याएँ पुनरुत्पन्न करना कठिन था। लेख कोने Science को corrected विश्लेषण pipeline और updated परिणम दिए हैं, जिनके बारे में उनका कहना है कि वे original से दशिसंख्यकीय महत्त्व और substantive आकार में मेल खाते हैं। Science उनका मूल्यंकन कर रहा है। जब तक वह प्रक्रिया पूरा नहीं होता, सटीक चित्र पर प्रश्न mark है जिसे केवल journal हटा सकता है।

इसलिए यहाँ एक सय दोकहमियाँ हैं: entrenched beliefs को तथ्य बदल सकते हैं यानही इस पर intriguing परिणम; और scientific रिकॉर्ड का खुद को खुले में सही करने का जिम्मा उदहरण। 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 प्रतर्भागीद्वारा खुद बताए सक्षम को rebut करने वाली तर्ज-दौर AI बातचीत ने चुनीषड्यंत्र पर मम्यता औसतन लगभग 20% घटाई; unrelated-topic नियंत्रण chat के विपरीत यह गरिष्ठ 10 days और 2 months बाद भी ममने योग्य थी। रिपोर्ट किया प्रभु durable लेकिन partial था अधिकिण प्रतर्भागीबद्ध में भीषड्यंत्र ममते थे — कमी cure नहीं। शोधपत्र पर फलिहल reporting inconsistencies और सखजनकि डेटसेट कीटुट्टि के कारण Editorial अभवियक्ति of Concern (Science, June 2026) है; लेखक कहते हैं corrected वशिलेण प्रभु कीदशि महत्त्व और आकर बनाए रखता है, जबकि Science अभी evaluate कर रहा है। यह चार्ट सफ शोधपत्र ने उसी सखजनकि डेटसेट से पुनर्नर्मित किया है, इसलिए दखाए मम अस्थयी हैं, शोधपत्र के अंतिम चित्र नहीं।

Original figure — The Clean Paper CC BY 4.0

शोधकर्तओं ने क्या किया

- 2190 प्रतर्भागीके सख दोप्रयोगे चलाए। हर प्रतर्भागीने अपने शब्दोंमें ऐसीषड्यंत्र सदिधंत बतई जिस पर वह वशिवस करताथा और वह सक्षम भीबतयाजो उसके हसिख से उसे समर्थन करताथा।
- हर प्रतर्भागीने GPT-4 Turbo के सख तीस-round बातचीत की। उपचार स्थितिमें AI को प्रतर्भागीके वशिष्ट सक्षम का rebuttal देने के लिए प्रमूट किया गया। नियंत्रण स्थितिमें उसने unrelated topic discuss किया।
- चुनीषड्यंत्र पर मम्यता बातचीत से पहले और तुरंत बाद ममा फिर प्रतर्भागीको 10 days और 2 months पर देखा संपर्क किया।
- Professional fact-checker से नमूनामें AI द्वारा किए गए सभी 128 factual दवे की सटीकता evaluate करई।
- जाँचा कि प्रभु दूसरी unrelated conspiracies तक spill over होता है या नहीं — और specificity परिक्षण के रूप में यह भी कि कहीं intervention उन conspiracies पर मम्यता तो नहीं घटता जो वस्तव में सत्य हैं।

क्यामलिया

- उपचार समूह में नियंत्रण की तुलना में चुनीषड्यंत्र पर मृत्युता में लगभग 20% औसत कमी (अध्ययन 1: 100-बटु पैमापर 95% विश्वास-अंतराल [13.8, 19.7] बटु, $P < 0.001$)।
- प्रभन्न कथन रहा। उपचार समूह में गरिष्ठ fade नहीं हुई: मृत्युता 10 days और दो months पर भी chat के तुरंत बटु के level के पास रही जबकि नियंत्रण अधिक रहा। शोधपत्र इसे चुनीषड्यंत्र पर कम-से-कम दो months तक undiminished प्रभन्न बतता है, momentary wobble नहीं।
- प्रभन्न अलग-अलग षड्यंत्र प्रकरोतक भी फैला और उन प्रतभिगी में भी रहाजनिकाप्रभन्निक मृत्युतामजबूत था।
- इस परस्थिति में AI के दत्ते accurate थे: 128 में 127 (99.2%) सत्य, 1 (0.8%) misleading, none असत्य।
- यह वशिष्ट था blanket doubt नहीं intervention ने सत्य conspiracies पर मृत्युता कम नहीं किया जिससे संकेत मलित है कि unsupported beliefs कदम हुई, लोग हर चीज के प्रति cynical नहीं बने।
- मध्यम spillover भी हुआ — other unrelated conspiracies पर मृत्युता लगभग 12% घटा और प्रतभिगी ने षड्यंत्र दत्ते का pushback करने की अधिक intention रिपोर्ट की। मुख्य प्रभन्न अध्ययन 2 में भी रहा और बनानियंत्रण समूह वत्ते छोटे नमूना में भी समान दिशा में गया (और कमजोर संक्षु, लेकिन संगत)।

यह क्या सति नहीं करता

- अभी यह बनावल वलापरणिम नहीं है। शोधपत्र data-handling और reproducibility समस्याओं के कारण Editorial अभव्यक्ति of Concern के तहत है, और corrected संख्याएँ कामूल्यंकन Science कर रहा है। समीक्षा पूरा होने तक वशिष्ट मत को अस्थायी मतों।
- यह AI षड्यंत्र मृत्युता “cure” करता है नहीं दिखाता ~20% औसत कमी साक्ष्य बदल है, erasure नहीं — अधिकिंश प्रतभिगी बटु में भी सदिधत मत थे, बस कम।
- यह वस्तुवकि दुनिया तैनीपरणिम नहीं है। प्रतभिगी अध्ययन में opt-in करके AI के साथ अच्छा faith में engaged हुए; वस्तुवकि दुनिया में षड्यंत्र के सबसे committed believers शयद उस chatbot को खोजने ही कम जाएँ जो उनसे तर्क देना करे।
- 99.2% सटीकता इस कह्य, मॉडल और सन्नधन prompting तक सीमित है — भणामॉडल के हमेशा सत्य हमें की समस्त्य गारंटी नहीं। लेखक स्पष्ट कहते हैं कि यही personalized persuasive शक्ति अगर असत्य beliefs की ओर लक्ष्य की जाए तो उन्हें भी बढा सकती है।
- “Durable” यहाँ दो months है, जो एकल बतचित के लिए impressive है, permanent के बरबर नहीं।

प्रमाण कतिनामजबूत है

- डजिइन वस्तुवकि strength है। नियंत्रति treatment-versus-नियंत्रण प्रयोग, साफ प्लेसबो-style नियंत्रण, दो अध्ययन और स्वतंत्र नमूना में प्रतकृति निर्माण, durability follow-ups, और AI के own दत्ते पर स्पष्ट सटीकता जाँच — persuasion शोध के बहुत से कम से यह अधिक सन्नधन है, और जहाँ जहाँ परिक्षण हुआ प्रभन्न की दिशा संगत रही।
- लेकिन अभव्यक्ति of Concern footnote नहीं है। Released डेटा और कोड से रिपोर्ट किया संख्याएँ regenerate कर पमापरणिम की trustworthiness का हिसा है, और यही चीज टूटी — जाँच-criteria inconsistencies और code-merging त्रुटि से duplicated/अतिरिक्त rows। लेखकों का कहना कि corrected pipeline परणिम बचती है encouraging और संभवति है, लेकिन फलिहल वह लेखकों का town account है, स्वतंत्र verdict नहीं। Science कामूल्यंकन नरिणयक

अगलाचरण है।

- **ईमजदार स्थिति “flagged” है, “debunked” या “पुष्ट” नहीं।** एक well-built, striking अध्ययन जिसके सटीक संख्याएँ औपचारिक समीक्षामें हैं। अच्छीकहमीकोयहीछोड़नाअसहज है, लेकिन फलिहल accurate स्थितियहीहै।

यह क्यों महत्वपूर्ण है

सलोतक एक प्रभमशलीview थाकि षड्यंत्र believers psychological needs और identity से driven होते हैं, इसलिए तथ्य से उन्हें मम्यतासे बहर तर्क देनानहीकियाजासकता। अगर durable सक्ष्य-based कमीटकिताहै, तोवह उस pessimism के खलिफ सस्थक counterweight है: शयद समस्याकाएक हसिसायह थाकि generic debunking कभीउस वशिष्ट सक्ष्य से engage नहींकरतीजसि believer वस्तव में cite करताहै — और LLM ऐसापैममापर कर सकताहै।

यह इस बत काममलाअध्ययन भीहै कि science कोकम कैसे करनाचहिए। अभव्यक्ति of Concern कासबक यह नहीं कि celebrated परणिम बेकर हैं; सबक यह है कि त्रुटियाँसतह होतीहैं, डेटाफरि examine होताहै, और दमे सखजनकि में re-check होते हैं। यह मशमरीचलनावशिषताहै, scandal नहीं — और इसीवजह से परणिम पर सहीposture applause या dismissal नहीं patience है।

AI के बारे में भीबत देमोदशिषों में जतीहै। Tailored तर्क से असत्य मम्यताकम करने कीवहीक्षमताकसीअसत्य मम्यताकोउतनीहीव्यग्रहकि रूप से बढ़ासकतीहै। Technique तटस्थ है; aim तटस्थ नहीं।

संक्षिप्त सर

2024 के एक अध्ययन ने पयाकि GPT-4 Turbo के सय तम-दरै बतची ने लोगोकीममीहुई षड्यंत्र सदिधंत पर मम्यता लगभग 20% घटई, प्रभम दोmonths तक रहाऔर अलग-अलग षड्यंत्र प्रकरोमें भीबनारहा AI के factual दमे overwhelming रूप से accurate grade किए गए। जून 2026 में data-handling और reproducibility समस्याएँ के कारण शेषपत्र पर Editorial अभव्यक्ति of Concern लगई गई; लेखक रिपोर्ट करते हैं कि corrected वशिषण नषिर्ष कीदशिा महत्व और आकर बनाए रखताहै, और Science अभीevaluate कर रहाहै। इसे वस्तव में दलिचस्प, carefully built परणिम कीतरह पदे जोफलिहल औपचारिक समीक्षामें हैं — 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>