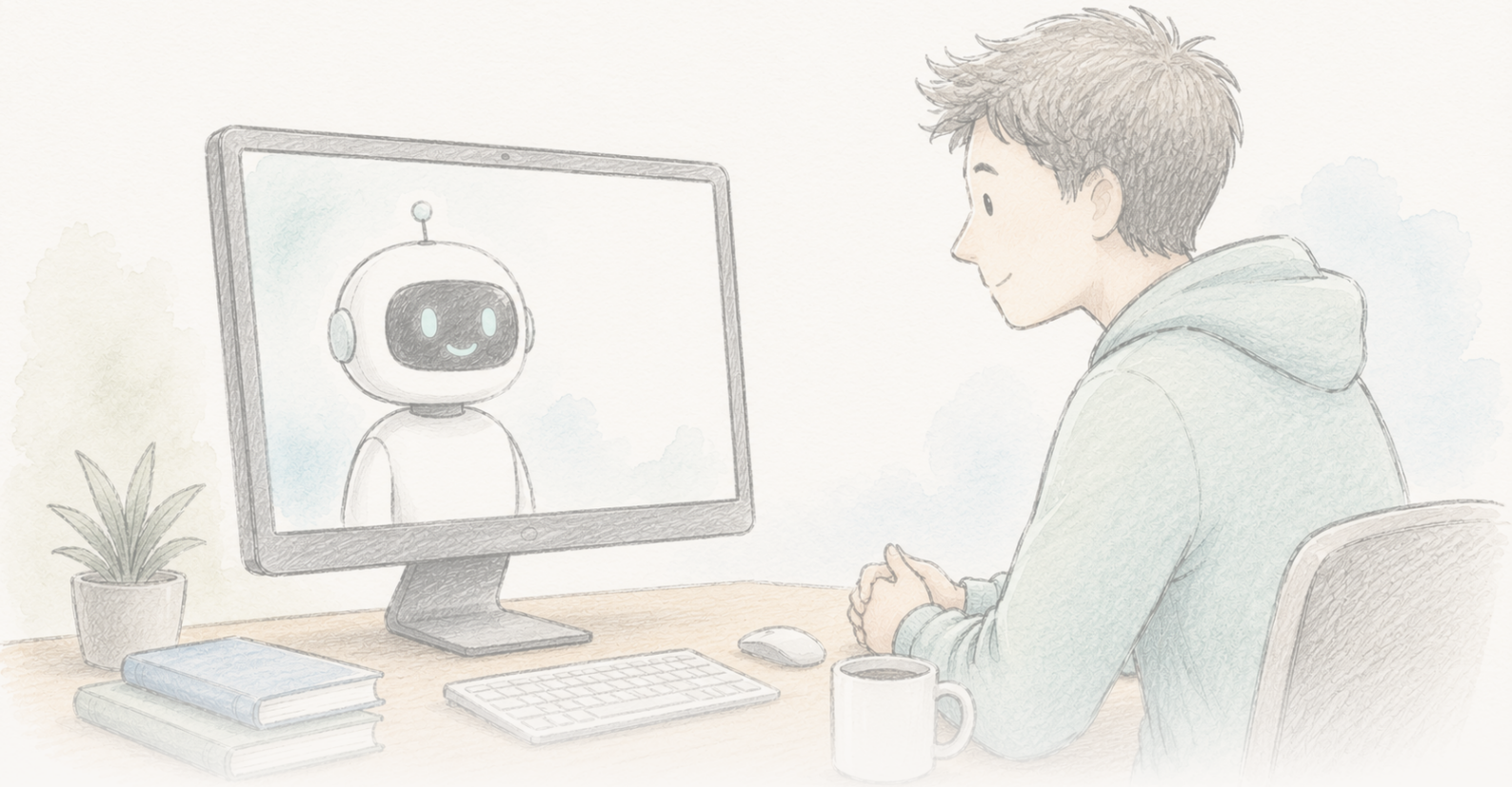


The CLEAN PAPER

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

یک chatbot ظاهراً باورهای توطئه‌ای را برای ماه‌ها کاهش داد

یک مطالعهٔ ۲۰۲۴ گزارش کرد گفت‌وگوی سه‌مرحله‌ای با *GPT-4 Turbo* باور افراد به نظریهٔ توطئه‌ای را که خودشان باور داشتند حدود ۲۰٪ کاهش داد؛ اثری که دو ماه ماند، در انواع توطئه دیده شد و بر ادعاهای *AI* با دقت بسیار بالا در همان تکلیف تکیه داشت. در ژوئن ۲۰۲۶ مقاله به‌دلیل مشکلات مدیریت داده و بازتولیدپذیری زیر *Expression Editorial Concern of* قرار گرفت؛ نویسندگان می گویند تحلیل اصلاح‌شده جهت، معناداری و اندازهٔ نتیجه را حفظ می کند و *Science* هنوز ارزیابی می کند. نتیجه‌ای واقعاً جالب و دقیق — اما فعلاً زیر بازبینی، نه تثبیت‌شده و نه ردشده.

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: science.adq1814/1126.10

Concern of Expression Editorial

Editorial Expression of Concern یک می‌کند، ارزیابی را بازتولیدپذیری و داده مدیریت به مربوط پرسش‌های که حالی در Science شود. تلقی موقت بررسی پایان تا باید نتیجه یعنی نیست؛ تخلف تشخیص یا retraction این است. افزوده مقاله این به Concern

→ Concern of Expression Editorial

بالاخره واقعیت‌ها — و یک پرچم هشدار روی مقاله

در سال ۲۰۲۴ مطالعه‌ای چیزی گزارش کرد که خلاف یک برداشت رایج و راحت بود. به کسی یک گفت‌وگوی کوتاه سه‌مرحله‌ای با AI بدهید — Turbo GPT-4 که طوری prompt شده تا به شواهد مشخصی پاسخ دهد که خود فرد برای نظریهٔ توطئه‌ای که شخصاً باور دارد مطرح می‌کند — و به‌طور میانگین باور او به آن نظریه حدود ۲۰٪ کم می‌شود. این کاهش دو ماه بعد هم باقی بود. اثر در توطئه‌های کلاسیک (ترور John F. Kennedy، بیگانگان، Illuminati و موضوعات روز، COVID-19، انتخابات ۲۰۲۰ آمریکا) دیده شد، حتی در کسانی که باورشان قوی و با هویتشان گره خورده بود. وقتی یک fact-checker حرفه‌ای ۱۲۸ ادعای واقعیتی AI را ارزیابی کرد، ۱۲۷ مورد درست، یک مورد گمراه‌کننده و هیچ کدام غلط نبود.

تیترِ همه‌گیر این بود که می‌شود با گفت‌وگو مردم را از تونل خرگوش بیرون آورد — باورمندان به توطئه خارج از دسترس شواهد نیستند؛ فقط به شواهد درست، با پاسخ‌گویی به‌اندازهٔ کافی شخصی نیاز دارند. این ادعایی واقعاً جالب است و مطالعه از بیشتر پژوهش‌های اقناع با دقت بیشتری ساخته شده بود.

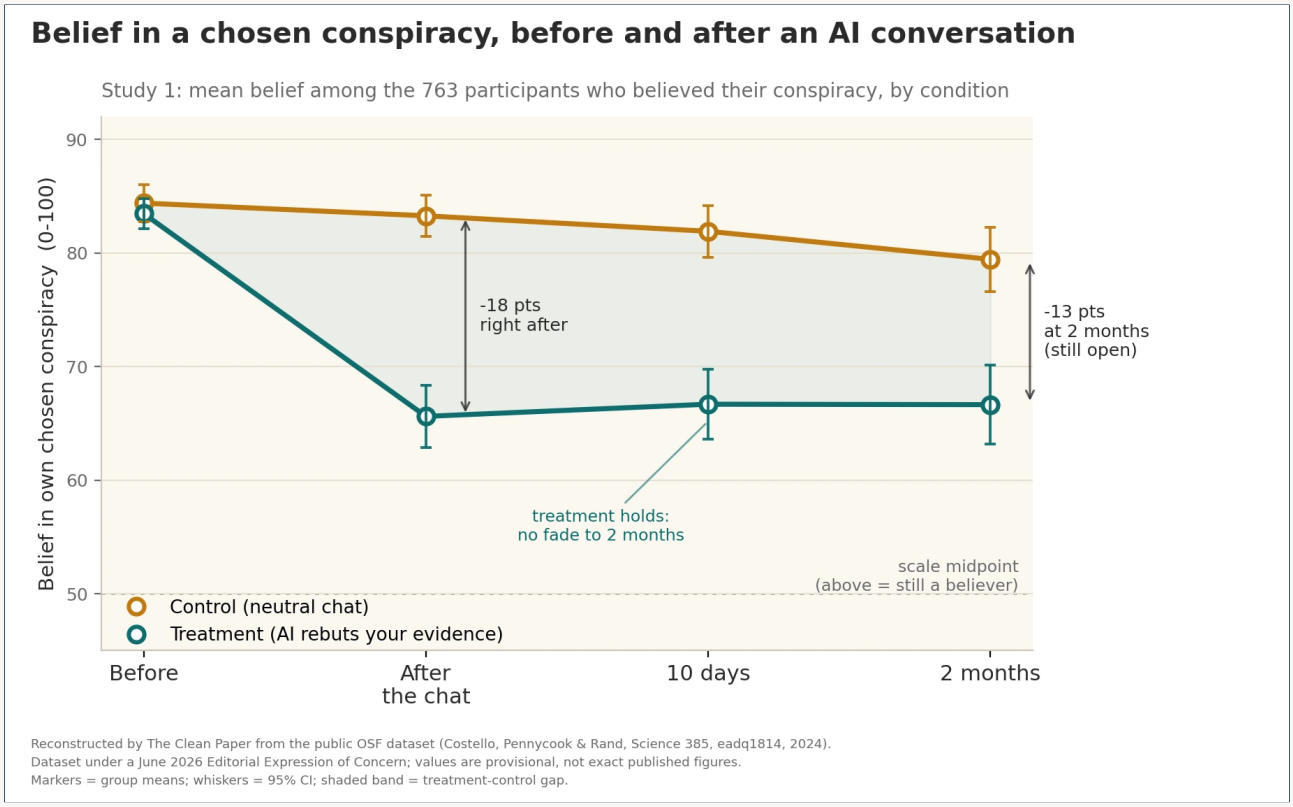
بعد، در ژوئن ۲۰۲۶، *Science* روی مقاله یک **Concern of Expression Editorial** گذاشت. این همان بخشی است که بیشتر بازگویی‌ها حذف خواهند کرد، و دقیقاً همان بخشی است که تعیین می‌کند فعلاً چقدر وزن می‌توان روی نتیجه گذاشت.

Concern of Expression چیست — و چیست نیست

پرسش‌هایی بگوید خوانندگان به تا می‌چسباند مقاله به مجله که است رسمی هشدار یک Editorial Expression of Concern تخلف تشخیص معنای به به‌خودی‌خود و نشده) گرفته پس (مقاله **نیست retraction** این بررسی‌اند. حال در هنوز و شده مطرح نویسندگان مورد، این در کند. تغییر است ممکن علمی سابقهٔ چون بخوانید، احتیاط این با است: این معنایش نیست. هم پژوهشی کردند. ارائه اصلاح‌شده تحلیل و دادند گزارش *Science* به را جزئیات کردند، بررسی مشکل‌ها از شدن آگاه از پس

مشکل‌های علامت‌گذاری‌شده مشخص‌اند. به نویسندگان اطلاع داده شد معیارهای غربالگری شرکت‌کنندگان میان متن مقاله و کد تحلیل منتشرشده به‌طور ناسازگار اعمال شده و دادهٔ عمومی ردیف‌های اضافه‌ای داشته که بر اثر خطای merge کد وارد شده بودند — مشکلاتی که بازتولید برخی عددهای گزارش‌شده را از مواد منتشرشده دشوار می‌کرد. نویسندگان یک pipeline اصلاح‌شده و مجموعه‌ای از نتایج به‌روز به *Science* دادند که می‌گویند از نظر **جهت، معناداری آماری و اندازهٔ عملی اثر** با اصل نتیجه مطابقت دارد. *Science* هنوز آن‌ها را ارزیابی می‌کند. تا پایان این ارزیابی، عددهای دقیق زیر علامت سؤالی هستند که فقط خود مجله می‌تواند بردارد.

پس این هم‌زمان دو داستان است: نتیجه‌ای جذاب دربارهٔ اینکه آیا واقعیت‌ها می‌توانند باورهای ریشه‌دار را جابه‌جا کنند، و نمونه‌ای زنده از اینکه رکورد علمی چگونه در فضای عمومی خودش را اصلاح می‌کند. هدف این نوشته جدا نگه داشتن این دو است.



در مطالعه گزارش شد یک گفت‌وگوی سه‌مرحله‌ای با AI که شواهد خودِ شرکت‌کننده برای نظریهٔ توطئهٔ منتخبش را رد می‌کرد، به‌طور میانگین باور او را حدود ۲۰٪ کاهش داد؛ برخلاف گفت‌وگوی کنترل دربارهٔ موضوعی بی‌ربط، این کاهش در ۱۰ روز و ۲ ماه بعد هنوز قابل‌اندازه‌گیری بود. اثر گزارش‌شده ماندگار اما ناقص بود: بیشتر شرکت‌کنندگان بعد از آن هنوز به توطئه باور داشتند — کاهش، نه درمان. مقاله اکنون به‌خاطر ناسازگاری‌های گزارش و خطا در دادهٔ عمومی زیر Concern of Expression Editorial (*Science*، ژوئن ۲۰۲۶) است؛ نویسندگان می‌گویند تحلیل اصلاح‌شده جهت، معناداری و اندازهٔ اثر را حفظ می‌کند، در حالی که *Science* هنوز ارزیابی را ادامه می‌دهد. این نمودار توسط The Clean Paper از همان دادهٔ عمومی بازسازی شده است، پس مقدارهای نشان‌داده‌شده موقت‌اند نه عددهای نهایی مقاله.

نویسندگان چه کردند

- دو آزمایش با ۲۱۹۰ شرکت‌کننده اجرا کردند؛ هر فرد با واژه‌های خودش یک نظریهٔ توطئه را که باور داشت و شواهدی را که فکر می‌کرد از آن حمایت می‌کنند نام برد.
- از هر شرکت‌کننده خواستند یک گفت‌وگوی سه‌مرحله‌ای با GPT-4 Turbo داشته باشد. در شرط **مداخله** AI طوری prompt شد که شواهد خاص فرد را رد کند؛ در شرط **کنترل** دربارهٔ موضوعی بی‌ربط صحبت کرد.
- باور به توطئهٔ منتخب را پیش از گفت‌وگو و بلافاصله بعد از آن اندازه گرفتند، سپس شرکت‌کنندگان را در **۱۰ روز و ۲ ماه** دوباره پیگیری کردند.
- از یک fact-checker حرفه‌ای خواستند در یک نمونه، صحت هر ۱۲۸ ادعای واقعیتی AI را ارزیابی کند.
- بررسی کردند آیا اثر به توطئه‌های بی‌ربط دیگر سرایت می‌کند و، به‌عنوان آزمون اختصاصی بودن، آیا باور به توطئه‌هایی را هم که واقعاً **درست** از آب درآمده‌اند کاهش می‌دهد یا نه.

چه پیدا کردند

- **کاهش میانگین حدود ۲۰٪** در باور به توطئهٔ منتخب در گروه مداخله نسبت به کنترل (مطالعهٔ ۱: فاصلهٔ اطمینان ۹۵٪ برابر [8.13, 7.19] واحد روی مقیاس ۱۰۰ امتیازی، P < 0.001).
- **اثر ماند.** در گروه مداخله کاهش محو نشد: باور در ۱۰ روز و دو ماه نزدیک سطح بلافاصله پس از گفت‌وگو باقی ماند و به سمت بالا برنگشت، در حالی که کنترل در تمام دوره بالاتر بود. مقاله اثر روی توطئهٔ منتخب را حداقل دو ماه بدون کاهش محسوس توصیف می‌کند، نه یک لغزش لحظه‌ای.
- **اثر تعمیم یافت** به دامنهٔ توطئه‌هایی که افراد نام بردند و حتی در شرکت‌کنندگانی که باور اولیه‌شان قوی بود دیده شد.
- **ادعاهای AI در این محیط دقیق بودند:** ۱۲۷ از ۱۲۸ (۹۹٫۲٪) درست، ۱ مورد (۰٫۸٪) گمراه‌کننده و هیچ موردی غلط ارزیابی شد.
- **اثر اختصاصی بود، نه بدبینی عمومی:** مداخله باور به توطئه‌های واقعی را **کاهش نداد**؛ نشانه‌ای اینکه باورهای بی‌پشتوانه را جابه‌جا کرد، نه اینکه افراد را نسبت به همه‌چیز بدبین کند.
- **سرایت اثر محدود اما قابل‌اندازه‌گیری بود:** باور به توطئه‌های دیگر و بی‌ربط حدود ۱۲٪ کاهش یافت و شرکت‌کنندگان قصد بیشتری برای مقابله با ادعاهای توطئه‌ای گزارش کردند. اثر اصلی در مطالعهٔ ۲ نیز پابرجا بود و در یک نمونهٔ کوچک‌ترِ بدون گروه کنترل هم همان جهت را نشان داد (شواهد ضعیف‌تر، اما سازگار).

این مقاله چه چیزی را اثبات نمی کند

- در حال حاضر **بی‌چالش نیست.** مقاله به‌خاطر مشکلات مدیریت داده و بازتولیدپذیری زیر Concern of Expression Editorial قرار دارد و عددهای اصلاح‌شده هنوز توسط *Science* ارزیابی می‌شوند. تا پایان آن بررسی، مقدارهای مشخص را موقت بدانید.
- نشان **نمی‌دهد** AI باور توطئه‌ای را «درمان» می‌کند. کاهش میانگین حدود ۲۰٪ تغییر معناداری است، نه پاک شدن باور — بیشتر افراد بعد از آن هنوز نظریه را باور داشتند، فقط کمتر.
- این **نتیجهٔ استقرار در دنیای واقعی نیست.** شرکت‌کنندگان داوطلب یک مطالعه بودند و با حسن نیت با AI وارد گفت‌وگو شدند؛ بیرون از آزمایش، کسانی که عمیق‌تر به یک توطئه متعهدند احتمالاً کمترین تمایل را دارند به یchatbot مراجعه کنند که با آن‌ها بحث کند.
- دقت ۹۹٫۲٪ **مختص همین کار، همین مدل و prompt دقیق** است — نه تضمینی عمومی که مدل‌های زبانی چیزهای درست می‌گویند. نویسندگان صریح‌اند که همین قدرت اقناع شخصی‌سازی‌شده اگر در جهت دیگری به کار رود می‌تواند باورهای غلط را هم تقویت کند.
- «ماندگار» اینجا یعنی **دو ماه**؛ برای یک گفت‌وگوی واحد قابل‌توجه است، اما به معنای دائمی نیست.

قدرت شواهد چقدر است؟

- **طراحی نقطهٔ قوت واقعی است.** آزمایش‌های کنترل‌شدهٔ مداخله در برابر کنترل، کنترل تمیزِ شبیه placebo، تکرار در دو مطالعه و یک نمونهٔ مستقل، پیگیری ماندگاری و بررسی صریح صحت ادعاهای خود AI — این از بیشتر پژوهش‌های اقناع دقیق‌تر است و جهت اثر هر جا آزموده شده سازگار است.
- **اما Concern of Expression پاورقی نیست.** اینکه بتوان عددهای گزارش‌شده را از داده و کد منتشرشده دوباره تولید کرد بخشی از اعتمادپذیری نتیجه است، و دقیقاً همین‌جا شکست رخ داده — ناسازگاری معیارهای غربالگری و ردیف‌های تکراری ناشی از خطای merge کد. گفتهٔ نویسندگان که pipeline اصلاح‌شده نتیجه را حفظ می‌کند امیدوارکننده و ممکن است درست باشد، اما فعلاً گزارش خود نویسندگان است نه داوری مستقل. ارزیابی *Science* چیزی است که باید منتظرش ماند.

- **وضعیت صادقانه «علامت گذاری‌شده» است، نه «ردشده» یا «تأییدشده».** مطالعه‌ای خوب‌طراحی‌شده و چشمگیر که عددهای دقیقش زیر بازبینی رسمی‌اند. شاید جای ناخوشایندی برای رها کردن یک داستان خوب باشد، اما جای دقیق همین است.

چرا مهم است

سال‌ها یک دیدگاه اثرگذار می‌گفت باورمندان به توطئه بیشتر توسط نیازهای روانی و هویت هدایت می‌شوند، پس با واقعیت نمی‌توان آن‌ها را از باورشان بیرون آورد. اگر کاهش پایدار و مبتنی بر شواهد این مطالعه پابرجا بماند، وزنه‌ای معنادار در برابر آن بدبینی است: شاید بخشی از مشکل این بوده که ردیه‌های عمومی هرگز سراغ شواهد خاصی که خود فرد به آن استناد می‌کند نمی‌رفتند، کاری که یک LLM می‌تواند در مقیاس بزرگ انجام دهد.

این مطالعه به‌عنوان نمونه‌ای از اینکه علم چگونه باید کار کند هم اهمیت دارد. درس Concern of Expression این نیست که نتیجه‌های مشهور بی‌ارزش‌اند؛ این است که خطاها آشکار می‌شوند، داده دوباره بررسی می‌شود و ادعاها در برابر چشم عموم دوباره آزموده می‌شوند. کار کردن این ماشین یک ویژگی است، نه رسوایی — و به همین دلیل موضع درست نسبت به این نتیجه صبر است، نه تشویق فوری و نه رد فوری.

و دربارهٔ AI دو لبه دارد. همان توانایی که می‌تواند با استدلالی متناسب یک باور غلط را ضعیف کند، می‌تواند به همان خوبی باوری غلط را باد کند. تکنیک خنثی است؛ هدف خنثی نیست.

خلاصهٔ تمیز

مطالعه‌ای در سال ۲۰۲۴ گزارش کرد گفت‌وگوی سه‌مرحله‌ای با Turbo GPT-4 باور افراد به یک نظریهٔ توطئه‌ای را که خودشان قبول داشتند حدود ۲۰٪ کاهش داد؛ اثر دو ماه ماند، در انواع توطئه‌ها دیده شد و ادعاهای واقعیتی AI در این محیط تقریباً همگی درست ارزیابی شدند. در ژوئن ۲۰۲۶ مقاله به‌دلیل مشکلات مدیریت داده و بازتولیدپذیری زیر Concern of Expression Editorial قرار گرفت؛ نویسندگان گزارش می‌کنند تحلیل اصلاح‌شده جهت، معناداری و اندازهٔ نتیجه را حفظ می‌کند و *Science* هنوز در حال ارزیابی است. آن را یک نتیجهٔ واقعاً جالب و خوب‌ساخته بخوانید که فعلاً زیر بازبینی است — نه تثبیت‌شده، نه ردشده. صادقانه‌ترین جمله دربارهٔ آن همان چیزی است که خود فرایند علمی می‌نویسد: دوباره بررسی‌اش کنید.

منابع

eadq1814 ,385 Science AI, with dialogues through beliefs conspiracy reducing Durably Rand, & Pennycook Costello, (2024)
https://doi.org/10.1126/science.adq1814

DOI Editor-in-Chief), Thorp, Holden H. ;2026 June 11 (posted Science Concern, of Expression Editorial science.aej2383/1126.10
https://www.science.org/doi/10.1126/science.aej2383