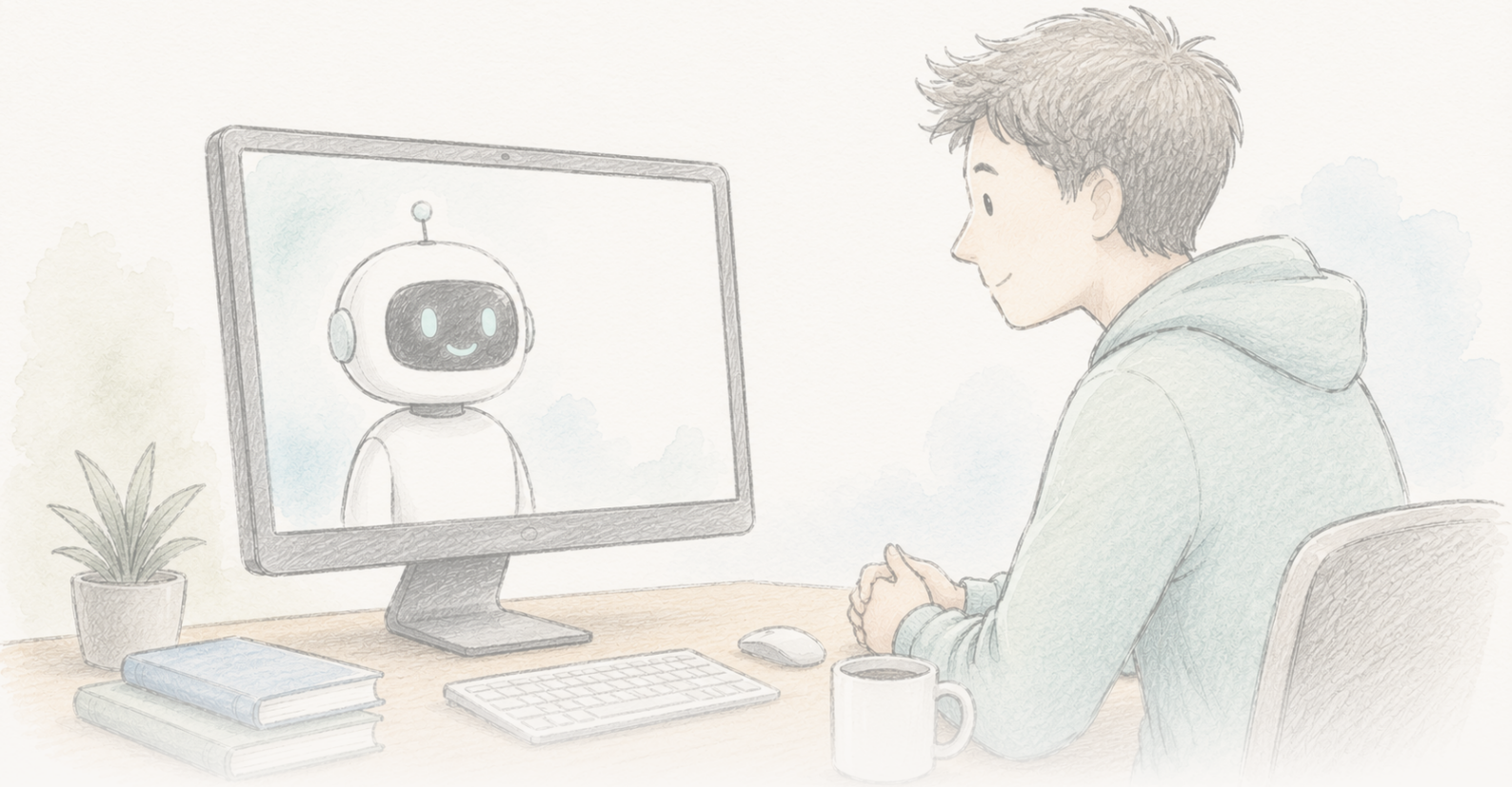


The Clean Paper

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

بدا أن روبوت دردشة يخفض الإيمان بنظريات المؤامرة لأشهر

وجدت دراسة عام 2024 أن محادثة من ثلاث جولات مع *GPT-4 Turbo* خفضت اعتقاد الناس في نظرية مؤامرة يؤمنون بها بنحو 20%، واستمر الأثر شهرين وظهر عبر أنواع مختلفة من المؤامرات، مع تقييم ادعاءات الذكاء الاصطناعي الواقعية بأنها دقيقة إلى حد ساحق. في يونيو 2026 وُضعت الورقة تحت *Concern of Expression Editorial* بسبب مشكلات في التعامل مع البيانات وقابلية الاستنساخ؛ ويقول المؤلفون إن تحليلاً مصححاً يحافظ على اتجاه النتيجة ودلالتها وحجمها، وما زالت *Science* تقيّمها. نتيجة مثيرة ومبنية بعناية — لكنها حالياً تحت المراجعة، لا محسومة ولا مدحوضة.

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

الاستنساخ. وقابلية البيانات مع بالتعامل تتعلق أسئلة تقيّم بينما Editorial Expression of Concern بالورقة Science ألحقت المراجعة. انتهاء حتى مؤقتة بوصفها قراءتها ينبغي المنشورة النتيجة أن يعني بل سلوك؛ سوء بوجود حكمًا ولا للورقة سحبًا ليس هذا

→ Concern of Expression Editorial

الحقائق، في نهاية المطاف — مع علامة تحذير على الورقة

في عام 2024، أبلغت دراسة عن نتيجة تخالف جزءًا مريحًا من الحكمة الشائعة. أعطِ شخصًا حوارًا قصيرًا من ثلاث جولات مع ذكاء اصطناعي — GPT-4 Turbo، مع توجيه يجعله يرد على الأدلة المحددة التي يستشهد بها الشخص لدعم نظرية مؤامرة يؤمن بها شخصيًا — وفي المتوسط يخفض إيمانه بتلك النظرية بنحو 20%. وظل الانخفاض موجودًا بعد شهرين. ظهر الأثر عبر مؤامرات كلاسيكية (اغتيال John F. Kennedy، والكائنات الفضائية، والمتنورون) وأخرى راهنة (COVID-19، وانتخابات الولايات المتحدة 2020)، وحتى لدى أشخاص كان إيمانهم قويًا ومرتبطًا بهويتهم. وعندما قيّم مدقق حقائق محترف 128 ادعاءً أطلقها الذكاء الاصطناعي، كان 127 منها صحيحًا، وواحدًا مضللًا، ولم يكن أي منها خاطئًا.

العنوان الذي انتشر كان أنك تستطيع إخراج الناس بالكلام من «بحر الأرنب» — أي أن المؤمنين بالمؤامرات ليسوا خارج متناول الأدلة، بل يحتاجون إلى الأدلة المناسبة المقدمة بدرجة كافية من التخصيص. هذا ادعاء مثير فعلًا، وكانت الدراسة مبنية بعناية أكبر من معظم أبحاث الإقناع.

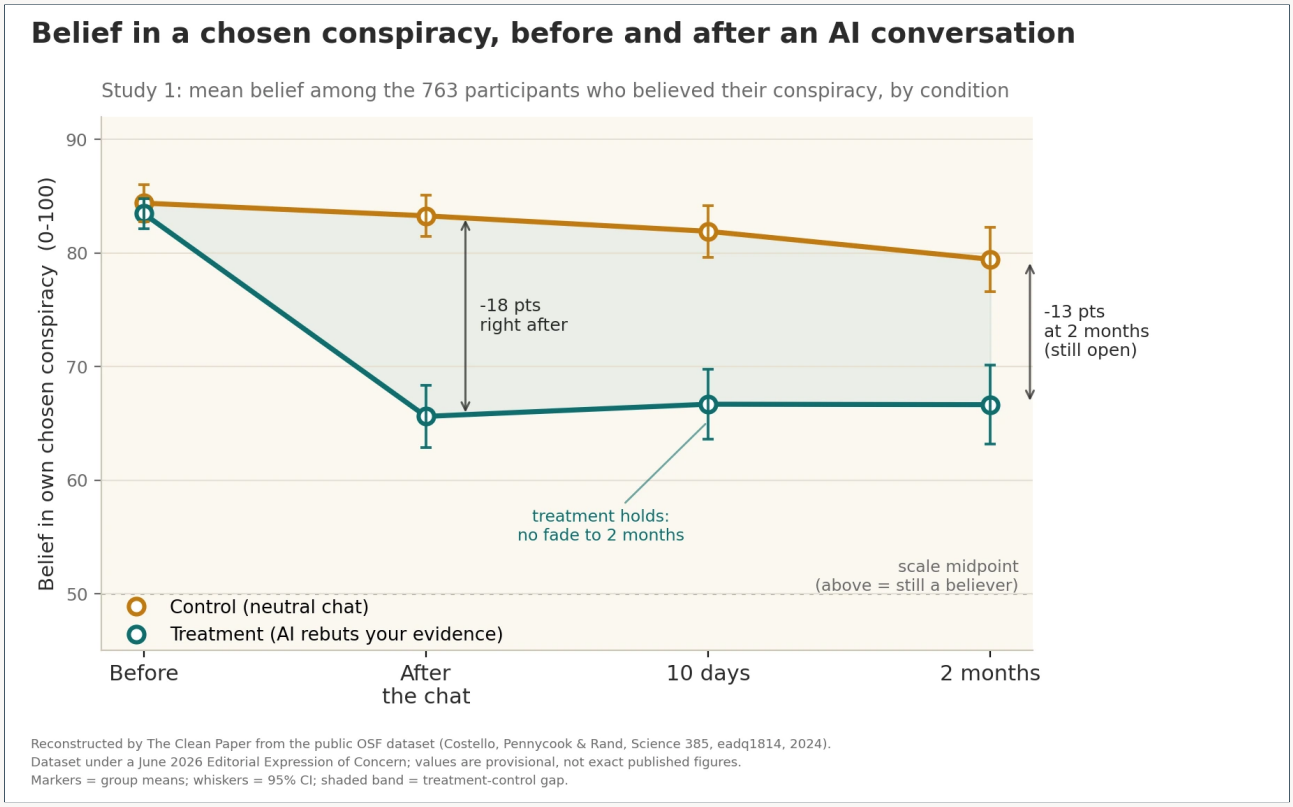
ثم، في يونيو 2026، نشرت *Science* **Concern of Expression Editorial** بشأن الورقة. وهذا هو الجزء الذي ستتجاوزه أغلب إعادة السرد، وهو بالتحديد الجزء الذي يحدد مقدار الوزن الذي يمكن أن تحمله النتيجة الآن.

ما هو إشعار هيئة التحرير بوجود دواعي للقلق — وما الذي لا يعنيه؟

القراء لإخبار بورقة مجلة تلحقه رسمي تنبيه هو (Editorial Expression of Concern) **للقلق دواعي بوجود التحرير هيئة إشعار** بوجود حكمًا ذاته حد في وليس تُسحب)، لم (فالورقة للورقة سحبًا **ليس** وهو جاريًا. فيها التحقيق زال ما بينما حولها أُثيرت أسئلة بأن *Science* وأبلغوا فيها، حققوا بالمشكلات، المؤلفون أُبلغ أن بعد هنا، يتغير. قد السجل لأن التحفظ هذا مع اقرأ معناه: سلوك. سوء مصححًا. تحليلًا وقدموا بالتفاصيل،

المشكلات التي أُشير إليها محددة. أُبلغ المؤلفون بوجود تناقضات في كيفية تطبيق معايير فرز المشاركين بين نص المخطوطة وكود التحليل المنشور، وبأن مجموعة البيانات العامة احتوت صفوفًا زائدة أُدرجت بسبب خطأ في دمج الكود — وهي مشكلات جعلت بعض الأرقام المنشورة صعبة الاستنساخ من المواد المتاحة. وقدّم المؤلفون إلى *Science* مسار تحليل مصححًا ومجموعة نتائج محدثة يقولون إنها تطابق الأصل **في الاتجاه والدلالة الإحصائية والحجم العملي للأثر.** وما زالت *Science* تقيّمها. وحتى ينتهي هذا التقييم، تبقى الأرقام الدقيقة تحت علامة استفهام لا تستطيع رفعها إلا المجلة.

إذن لدينا قصتان في الوقت نفسه: نتيجة مثيرة بشأن ما إذا كانت الحقائق قادرة على تحريك معتقدات راسخة، ومثال حي على أن السجل العلمي يصحح نفسه علنًا. هدف هذا المقال هو إبقاء القصتين منفصلتين بوضوح.



في الدراسة، أُبلغ أن محادثة من ثلاث جولات مع الذكاء الاصطناعي، ترد على الأدلة التي ذكرها المشارك نفسه، خفضت اعتقاده في المؤامرة التي اختارها بنحو 20% في المتوسط، وعلى خلاف محادثة ضابطة في موضوع غير ذي صلة، ظل الانخفاض قابلًا للقياس بعد 10 أيام وشهرين. كان الأثر المبلغ عنه مستمرًا لكنه جزئيًا: ظل معظم المشاركين يؤمنون بالمؤامرة بعد ذلك — انخفاضًا لا «علاجًا». الورقة تخضع حاليًا لإشعار من هذا النوع صادر عن *Science* (يونيو 2026) بسبب تناقضات في الإبلاغ وخطأ في مجموعة البيانات العامة؛ ويقول المؤلفون إن التحليل المصحح يحافظ على اتجاه الأثر ودلالته وحجمه، بينما تواصل *Science* التقييم. أعادت Paper Clean The بناء هذا الرسم من مجموعة البيانات العامة، لذلك القيم المعروضة مؤقتة وليست الأرقام النهائية للورقة.

Original figure — The Clean Paper CC BY 0.4

ما الذي فعله الباحثون

- أجروا تجربتين مع 2190 مشاركًا، سمّى كل واحد منهم — بكلماته الخاصة — نظرية مؤامرة يؤمن بها، والأدلة التي يعتقد أنها تدعمها.
- أجرى كل مشارك محادثة من ثلاث جولات مع GPT-4 Turbo. في حالة **التدخل** وُجّه الذكاء الاصطناعي إلى تفنيد الأدلة المحددة التي ذكرها المشارك؛ وفي حالة **الضبط** ناقش موضوعًا غير ذي صلة.
- قاسوا الاعتقاد في المؤامرة المختارة قبل المحادثة وبعدها مباشرة، ثم أعادوا الاتصال بالمشاركين بعد **10 أيام وشهرين**.
- طلبوا من مدققي حقائق محترف تقييم دقة جميع الادعاءات الواقعية الـ128 التي قدمها الذكاء الاصطناعي في عينة.
- اختبروا ما إذا كان الأثر يمتد إلى مؤامرات أخرى غير مرتبطة — وكاختبار للخصوصية، ما إذا كان يخفض أيضًا الاعتقاد في مؤامرات ثبت أنها **حقيقية**.

ماذا وجدوا

- **انخفاضًا متوسطه نحو 20%** في الاعتقاد بالمؤامرة المختارة لدى مجموعة التدخل مقارنة بالضبط (الدراسة 1: فاصل ثقة 95% [8.13، 7.19] نقطة على مقياس من 100، P < 0.001).
- **استمر الأثر.** في مجموعة التدخل لم يتلاش الانخفاض: ظل الاعتقاد قريبًا من مستواه بعد المحادثة عند 10 أيام وشهرين بدل أن يرتفع تدريجيًا، بينما ظلت مجموعة الضبط أعلى طوال الوقت. تصف الورقة الأثر على المؤامرة المختارة بأنه استمر بلا ضعف ملحوظ لمدة شهرين على الأقل، لا كتذبذب لحظي.
- **كان عامًا عبر الأنواع** المختلفة من المؤامرات التي ذكرها المشاركون، واستمر حتى لدى من كان اعتقادهم قويًا في البداية.
- **كانت ادعاءات الذكاء الاصطناعي دقيقة** في هذا السياق: 127 من 128 (99.2%) قُيّمت صحيحة، وواحد (0.8%) مضللًا، ولا شيء خاطئًا.
- **كان الأثر محددًا** لا شكًا عامًا: لم يخفض التدخل الاعتقاد في مؤامرات حقيقية، ما يوحي بأنه حرّك المعتقدات غير المدعومة بدل جعل الناس ساخرين من كل شيء.
- **امتد بدرجة متواضعة** إلى مؤامرات أخرى غير مرتبطة — انخفض الاعتقاد فيها بنحو 12%، وأبلغ المشاركون عن نية أكبر لمعارضة الادعاءات التآمرية. وصمد الأثر الرئيسي في الدراسة الثانية، واتجه في الاتجاه نفسه في عينة أصغر بلا مجموعة ضبط (دليل أضعف، لكنه متسق).

ما الذي لا تثبته الدراسة

- لا تقف النتيجة **من دون أسئلة في الوقت الحالي.** الورقة تحت Concern of Expression Editorial بسبب مشكلات في التعامل مع البيانات وقابلية الاستنساخ، والأرقام المصححة ما زالت قيد تقييم *Science*. تعامل مع القيم المحددة بوصفها مؤقتة حتى ينتهي هذا التقييم.
- لا تُظهر أن الذكاء الاصطناعي «يعالج» الإيمان بالمؤامرات. انخفاض متوسطه نحو 20% تحول ذو معنى، لا محوًا للاعتقاد — ظل معظم المشاركين مؤمنين بالنظرية بعد ذلك، لكن بدرجة أقل.
- هذه **ليست** نتيجة نشر في العالم الحقيقي. المشاركون اختاروا الدخول في دراسة وتفاعلوا مع الذكاء الاصطناعي بحسن نية؛ خارج المختبر، قد يكون أكثر المؤمنين بالمؤامرات تشددًا هم الأقل احتمالًا لطلب روبوت دردشة يجادلهم.
- دقة 99.2% **خاصة بهذه المهمة وهذا النموذج وهذا التوجيه الحذر** — وليست ضمانًا عامًا بأن نماذج اللغة تقول الحقيقة. والمؤلفون واضحون في أن القدرة الإقناعية الشخصية نفسها يمكن أن تدفع معتقدات خاطئة إذا وُجهت لذلك.
- «مستمر» هنا يعني **شهرين**، وهو أمر لافت لمحادثة واحدة، لكنه ليس مرادفًا للديمومة.

ما مدى قوة الدليل

- **التصميم نقطة قوة حقيقية.** تجارب مضبوطة تدخل مقابل ضبط، وضبط نظيف شبيه بالدواء الوهمي، وتكرار عبر دراستين وعينة مستقلة، ومتابعات للاستمرارية، وفحص صريح لدقة ادعاءات الذكاء الاصطناعي نفسه — كل ذلك أكثر عناية من معظم أبحاث الإقناع، واتجاه الأثر متسق حيثما اختُبر.
- **لكن إشعار هيئة التحرير بوجود دواعي للقلق ليس هامشًا صغيرًا.** القدرة على إعادة توليد الأرقام المنشورة من البيانات والكود المتاحين جزء من الثقة في النتيجة، وهذا تحديدًا ما تعطّل: تناقضات في معايير الفرز وصفوف مكررة بسبب خطأ في دمج الكود. قول المؤلفين إن المسار المصحح يحافظ على النتيجة مشجع ومعقول، لكنه يظل تقرير المؤلفين أنفسهم وليس حكمًا مستقلًا بعد. تقييم *Science* هو ما ينبغي

انتظاره.

- **الحالة الصادقة هي «معلَّمة بملاحظة»، لا «دُحضت» ولا «تأكدت».** دراسة قوية التصميم ولافتة، لكن أرقامها الدقيقة تحت مراجعة رسمية. قد يكون ذلك مكانًا غير مريح لترك قصة جيدة، لكنه المكان الدقيق.

لماذا يهم ذلك

لسنوات، كان هناك رأي مؤثر يقول إن المؤمنين بنظريات المؤامرة تحركهم احتياجات نفسية وهوية، ولذلك لا يمكن إخراجهم من معتقداتهم بالحقائق. إذا صمد انخفاض مستمر قائم على الأدلة، فهو ثقل حقيقي في مواجهة هذا التشاؤم: ربما كانت المشكلة جزئيًا أن التفنيد العام لا يتعامل أبدًا مع الدليل المحدد الذي يستشهد به المؤمن فعلًا، وهو شيء يستطيع نموذج لغوي كبير فعله على نطاق واسع.

كما يهم ذلك بوصفه دراسة حالة في الطريقة التي يفترض أن يعمل بها العلم. والدرس من هذا الإشعار ليس أن النتائج الشهيرة عديمة القيمة؛ بل أن الأخطاء تظهر، والبيانات يعاد فحصها، والادعاءات يعاد اختبارها علنًا. عمل هذه الآلية ميزة لا فضيحة — ولهذا الموقف المناسب من النتيجة هو الصبر، لا التصفيق ولا الرفض.

والأمر ذو حدين في الذكاء الاصطناعي. القدرة نفسها على إضعاف معتقد خاطئ بحجة مصممة لشخص بعينه قد تستطيع تضخيم معتقد خاطئ بالفعالية نفسها. التقنية محايدة؛ الهدف ليس كذلك.

خلاصة واضحة

وجدت دراسة عام 2024 أن محادثة من ثلاث جولات مع GPT-4 Turbo خفضت اعتقاد الناس في نظرية مؤامرة يؤمنون بها بنحو 20%، مع أثر استمر شهرين وظهر عبر أنواع متعددة من المؤامرات، بينما قُيّمت ادعاءات الذكاء الاصطناعي الواقعية بأنها دقيقة إلى حد ساحق. في يونيو 2026 وُضعت الورقة تحت Concern of Expression Editorial بسبب مشكلات في التعامل مع البيانات وقابلية الاستنساخ؛ ويقول المؤلفون إن تحليلًا مصححًا يحافظ على اتجاه النتيجة ودلالتها وحجمها، وما زالت *Science* تقيّم ذلك. اقرأها بوصفها نتيجة مثيرة فعلًا ومبنية بعناية لكنها قيد المراجعة حاليًا — لم تُحسم ولم تُدحض. أكثر جملة أمينة بشأنها هي التي تكتبها العملية العلمية نفسها: افحصها مرة أخرى.

المصادر

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