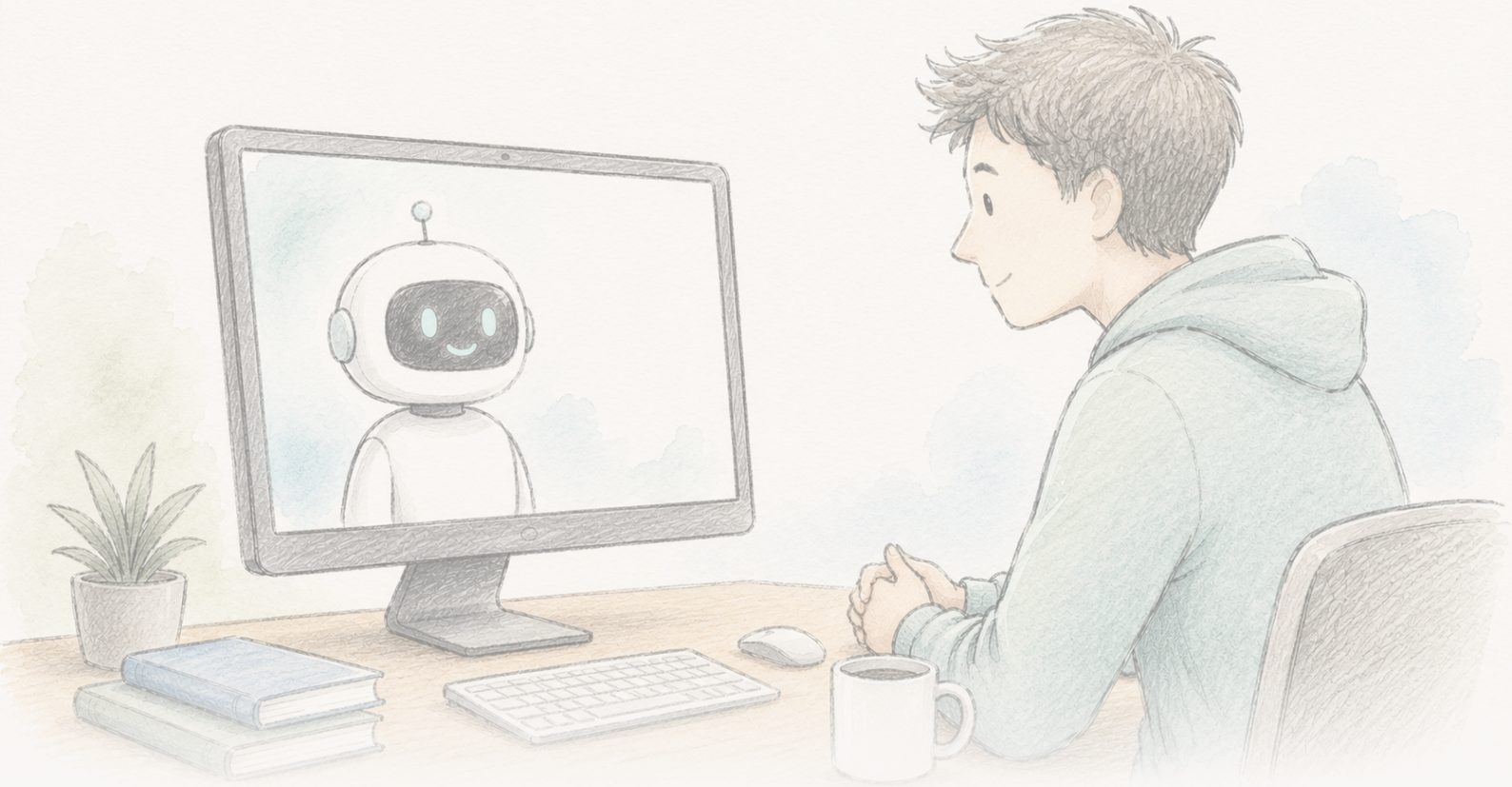


The CLEAN PAPER

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

צ'אטבוט נראה כמפחית אמונות קונספירטיביות למשך חודשים

מחקר מ-*2024* מצא ששיחה בת שלושה סבבים עם *GPT-4 Turbo* הפחיתה בכ-*20%* את אמונתם של אנשים בתאוריית קונספירציה שבה החזיקו, אפקט שנמשך חודשים והופיע בסוגי קונספירציות שונים, כאשר טענות ה-*AI* דורגו כמעט כולן כנכונות. ביוני *2026* הוצמד למאמר *Concern of Expression Editorial* בשל בעיות טיפול בנתונים ושחזור, המחברים אומרים שניתוח מתוקן משמר את הממצא בכיוון, במובהקות ובגודל, ו-*Science* עדיין בוחן אותו. תוצאה מעניינת באמת ובנויה בזהירות — אך כרגע תחת בדיקה, לא סגורה ולא מופרכת.

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

Concern of Expression Editorial

בנתונים לטיפול בנוגע שאלות בוחן שהוא בזמן Editorial Expression of Concern למאמר הצמיד Science התוצאה את לקרוא שיש היא המשמעות הולמת; בלתי התנהגות של קביעה ואינה רטרקציה אינה זו ולשחזור. הבדיקה. לסיום עד כזמנית

→ Concern of Expression Editorial

בסופו של דבר, עובדות — ודגל אזהרה על המאמר

ב-2024 דיווח מחקר על משהו שסותר תפיסה נוחה ונפוצה. תנו לאדם שיחה קצרה בת שלושה סבבים עם AI — Turbo, GPT-4 שהונחה לענות לראיות הספציפיות שהוא מציג לטובת תאוריית קונספירציה שבה הוא עצמו מאמין — ובממוצע אמונתו באותה תאוריה יורדת בכ-20%. הירידה עדיין הייתה נוכחת חודשיים לאחר מכן. זה עבד על קונספירציות קלאסיות (רצח Kennedy, F. John חייזרים, האילומינטי) וגם על אקטואליות (COVID-19, הבחירות בארה"ב ב-2020), ואפילו אצל אנשים שאמונתם הייתה חזקה וקשורה לזהותם. כאשר בודק עובדות מקצועי דירג 128 טענות שה-AI העלה, 127 נמצאו נכונות, אחת מטעה ואף אחת לא שקרית.

הכותרת שהתפשטה הייתה שאפשר בכל זאת לדבר אנשים החוצה מ"מחילת הארנב" — שמאמיני קונספירציות אינם מחוץ להישג ידן של ראיות; הם פשוט צריכים את הראיות הנכונות, המותאמות לטענות שלהם. זו טענה מעניינת באמת, והמחקר נבנה בזהירות רבה יותר מרוב עבודות השכנוע.

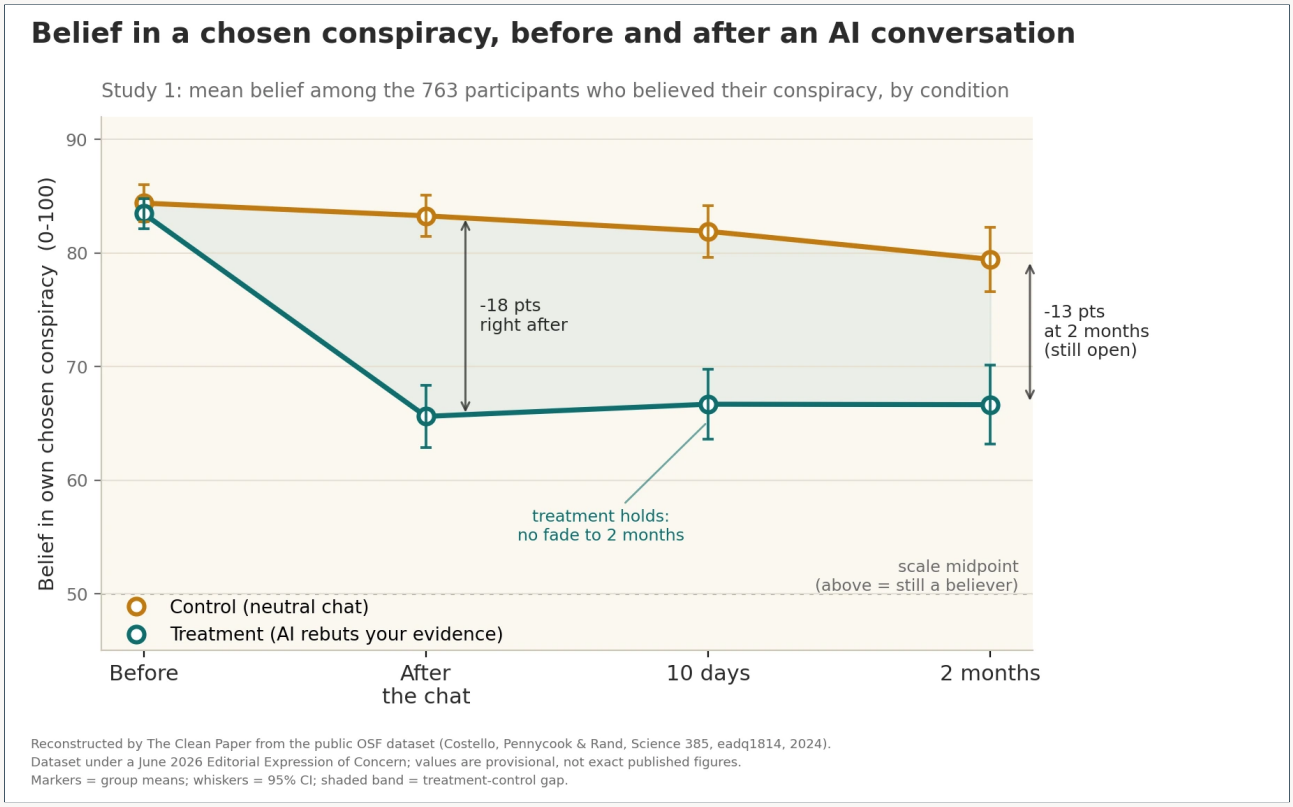
ואז, ביוני 2026, *Science* פרסם על המאמר **Concern of Expression Editorial**. זה החלק שרוב הסיפורים החוזרים ידלגו עליו, וזה בדיוק החלק שקובע כמה משקל אפשר לתת לתוצאה כרגע.

מהו Concern of Expression — ומה הוא אינו

שעלו קוראים להזהיר כדי למאמר מצמיד שכתב-עת רשמי סימון הוא Editorial Expression of Concern התנהגות של קביעה כשלעצמה אינה והיא נמשך), לא (המאמר רטרקציה **אינה** זו בבדיקה. שעדיין שאלות עשויה עוד המדעית שהרשומה מפני הזאת, ההסתייגות עם המאמר את קראו היא: המשמעות הולמת. בלתי הפרטים על *Science* ל- דיווחו אותן, חקרו המחברים לידיעתם, הובאו שהבעיות לאחר הזה, במקרה להשתנות. מתוקן. ניתוח וסיפקו

הדברים שסומנו ספציפיים. למחברים נודע על אי-עקביות בין קריטריוני סינון המשתתפים שתוארו בכתב היד לבין אלה שיושמו בקוד הניתוח שפורסם, וכן שמערך הנתונים הציבורי כלל שורות מיותרות ששולבו בטעות עקב שגיאת מיזוג קוד — בעיות שהקשו לשחזר חלק מהמספרים שדווחו מתוך החומרים ששוחררו. המחברים מסרו ל-*Science* pipeline מתוקן לניתוח וקבוצת תוצאות מעודכנת, שלדבריהם תואמת את המקור **בכיוון, במובהקות הסטטיסטית ובגודל המהותי**. *Science* עדיין מעריך אותן. עד שההערכה תסתיים, המספרים המדויקים נושאים סימן שאלה שרק כתב-העת יכול להסיר.

לכן יש כאן שני סיפורים בו-זמנית: תוצאה מסקרנת על השאלה אם עובדות יכולות להזיז אמונות מושרשות, ודוגמה חיה לאופן שבו הרשומה המדעית מתקנת את עצמה בפומבי. מטרת הקטע הזה היא לא לערבב ביניהם.



במחקר דווח ששיחה בת שלושה סבבים עם AI, שסתרה את הראיות שהמשתתף עצמו הציג, הפחיתה את האמונה בקונספירציה שבחר בכ-20% בממוצע; בניגוד לשיחת ביקורת בנושא לא קשור, הירידה עדיין הייתה ניתנת למדידה לאחר 10 ימים וחודשיים. האפקט המדווח היה מתמשך אך חלקי: רוב המשתתפים עדיין האמינו בקונספירציה לאחר מכן — הפחתה, לא "ריפוי". המאמר נמצא כעת תחת Concern of Expression Editorial (*Science*, יוני 2026) בשל אי-עקביות בדיווח ושגיאה במערך הנתונים הציבורי; המחברים אומרים שניתוח מתוקן שומר על כיוון האפקט, מובהקותו וגודלו, בזמן ש-*Science* ממשיך להעריך. הגרף שוחזר ב-The Clean Paper מתוך מערך הנתונים הציבורי, ולכן הערכים המוצגים הם זמניים ולא המספרים הסופיים של המאמר.

מה עשו המחברים

- ערכו שני ניסויים עם 2,190 משתתפים, שכל אחד מהם ציין במילותיו תאוריית קונספירציה שבה הוא מאמין ואת הראיות שלדעתו תומכות בה.
- כל משתתף ניהל שיחה בת שלושה סבבים עם GPT-4 Turbo. בתנאי **הטיפול** ה-AI הונחה לסתור את הראיות הספציפיות של המשתתף; בתנאי **הביקורת** הוא דן בנושא לא קשור.
- מדדו את האמונה בקונספירציה שנבחרה לפני השיחה ומיד אחריה, ואז חזרו למשתתפים לאחר 10 **ימים וחודשיים**.
- ביקשו מבודק עובדות מקצועי להעריך את דיוק כל 128 הטענות העובדתיות שה-AI העלה במדגם.
- בדקו אם האפקט "נשפך" גם לקונספירציות אחרות, לא קשורות — וכבדיקת ספציפיות, אם הוא גם מוריד אמונה בקונספירציות שבמקרה הן **נכונות**.

מה הם מצאו

- **ירידה ממוצעת של בערך 20%** באמונה בקונספירציה שנבחרה בקבוצת הטיפול יחסית לביקורת (מחקר 1: רווח סמך 95% של [13.8, 19.7] נקודות בסולם של 100 נקודות, P < 0.001).
- **היא נמשכה.** בקבוצת הטיפול הירידה לא נעלמה: האמונה נשארה קרובה לרמתה מיד לאחר השיחה גם לאחר 10 ימים וחודשיים, במקום לטפס חזרה, בעוד הביקורת נשארה גבוהה יותר לאורך כל התקופה. המאמר מתאר את האפקט על הקונספירציה שנבחרה כנמשך ללא היחלשות לפחות חודשיים, לא כטלטלה רגעית.
- **האפקט הכליל** על פני מגוון הקונספירציות שהמשתתפים ציינו, ונשמר גם אצל מי שאמונתו בתחילת המחקר הייתה חזקה.
- **טענות ה-AI היו מדויקות** בהקשר הזה: 127 מתוך 128 (99.2%) דורגו כנכונות, 1 (0.8%) כמטעה ואף אחת כשקרית.
- **האפקט היה ספציפי**, לא ספקנות כללית: ההתערבות **לא** הפחיתה אמונה בקונספירציות אמיתיות, מה שמרמז שהיא הזיזה אמונות חסרות ביסוס במקום לגרום לאנשים לפקפק בהכול.
- **הייתה גם זליגה מתונה** — האמונה בקונספירציות אחרות, לא קשורות, ירדה בכ-12%, והמשתתפים דיווחו על כוונה גדולה יותר להתנגד לטענות קונספירטיביות. האפקט המרכזי נשמר במחקר 2 והצביע לאותו כיוון במדגם קטן יותר ללא קבוצת ביקורת (ראיה חלשה יותר, אך עקבית).

מה זה לא מוכיח

- התוצאה **אינה עומדת כרגע ללא סימן שאלה.** המאמר תחת Concern of Expression Editorial בשל בעיות טיפול בנתונים ושחזור, והמספרים המתוקנים עדיין בהערכה של *Science*. יש להתייחס לערכים הספציפיים כזמניים עד שהבדיקה תסתיים.
- זה **לא** מראה ש-AI "מרפא" אמונה בקונספירציות. ירידה ממוצעת של ~20% היא שינוי משמעותי, לא מחיקה — רוב המשתתפים עדיין האמינו בתאוריה לאחר מכן, רק פחות.
- זו **אינה** תוצאת פריסה בעולם האמיתי. המשתתפים הסכימו להשתתף במחקר וניהלו את השיחה בתום לב; בעולם הפתוח, האנשים המחויבים ביותר לקונספירציה עשויים להיות דווקא הפחות מוכנים לפנות לצ'אטבוט שיתווכח איתם.
- דיוק של 99.2% **ספציפי למשימה, למודל ול-prompting הזהיר** — לא ערובה כללית לכך שמודלי שפה אומרים אמת. המחברים מציינים במפורש שאותה יכולת שכנוע מותאמת אישית יכולה לקדם גם אמונות שקריות אם תכוון לשם.
- "מתמשך" כאן פירושו **חודשיים** — מרשים לשיחה יחידה, אך לא "קבוע".

עד כמה הראיות חזקות

- **התכנון הוא חוזקה אמיתית.** ניסויים מבוקרים של טיפול מול ביקורת, ביקורת דמוית-פלצבו נקייה, שכפול בשני מחקרים ובמדגם עצמאי, מעקב לאורך זמן ובדיקת דיוק מפורשת לטענות ה-AI — זה זהיר יותר מרוב מחקרי השכנוע, וכיוון האפקט עקבי בכל המקומות שבהם נבדק.
- **אבל** Concern of Expression **אינו הערת שוליים.** היכולת להפיק מחדש את המספרים המדווחים מן הנתונים והקוד המשוחררים היא חלק מאמינות התוצאה, וזה בדיוק מה שנשבר — אי-עקביות בקריטריוני הסינון ושורות משוכפלות/מיותרות עקב שגיאת מיזוג קוד. הצהרת המחברים ש-pipeline מתוקן שומר על התוצאה מעודדת

וסבירה, אבל זו עדיין הצהרת המחברים, לא פסק דין עצמאי. ההערכה של *Science* היא הדבר שצריך להמתין לו.

- **הסטטוס הכנה הוא "מסומן לבדיקה", לא "הופרך" ולא "אושר".** מחקר בנוי היטב ומרשים, שהמספרים המדויקים שלו נמצאים בבדיקה רשמית. זה מקום לא נוח להשאיר בו סיפור טוב — והוא המקום המדויק.

למה זה חשוב

במשך שנים הייתה השפעה לתפיסה שלפיה מאמיני קונספירציות מונעים מצרכים פסיכולוגיים וזהות, ולכן אי-אפשר לשכנע אותם לצאת מאמונתם באמצעות עובדות. אם היא תחזיק מעמד, ירידה מתמשכת המבוססת על ראיות תהיה משקל נגד משמעותי לפסימיות הזאת: היא מרמזת שחלק מהבעיה היה שהפרכות כלליות מעולם לא נגעו ב-ראיה הספציפית שמאמין מסוים מצטט — דבר ש-LLM יכול לעשות בקנה מידה גדול.

המחקר חשוב גם כמקרה מבחן לאופן שבו המדע אמור לעבוד. הלקח מ-Concern of Expression אינו שתוצאות מפורסמות חסרות ערך; אלא ששגיאות נחשפות, נתונים נבדקים מחדש וטענות נבחנות שוב בפומבי. העובדה שהמנגנון הזה פועל היא תכונה, לא שערורייה — והיא הסיבה שהעמדה הנכונה כלפי התוצאה כרגע היא סבלנות, לא מחיאות כפיים ולא ביטול.

וזה פועל לשני הכיוונים גם לגבי AI. אותה יכולת להחליש אמונה שקרית בעזרת טיעון מותאם יכולה באותה מידה לחזק אחת. הטכניקה נייטרלית; המטרה אינה.

בקצרה

מחקר מ-2024 מצא ששיחה בת שלושה סבבים עם GPT-4 Turbo הפחיתה בכ-20% את אמונתם של אנשים בתאוריית קונספירציה שבה האמינו, אפקט שנמשך חודשיים והכליל על פני סוגי קונספירציות, כאשר הטענות העובדתיות של ה-AI דורגו כמעט כולן כנכונות. ביוני 2026 הוצמד למאמר Concern of Expression Editorial בשל בעיות טיפול בנתונים ושחזור; המחברים מדווחים שניתוח מתוקן משמר את הממצא בכיוון, במובהקות ובגודל, ו-*Science* עדיין מעריך אותו. נכון לקרוא את המחקר כתוצאה מעניינת באמת ובנויה בזהירות שנמצאת כרגע תחת בדיקה — לא סגורה, ולא מופרכת. המשפט הכנה ביותר לגביו הוא זה שהתהליך עצמו כותב: לבדוק שוב.

מקורות

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 10.1126/science.aej2383
https://www.science.org/doi/10.1126/science.aej2383