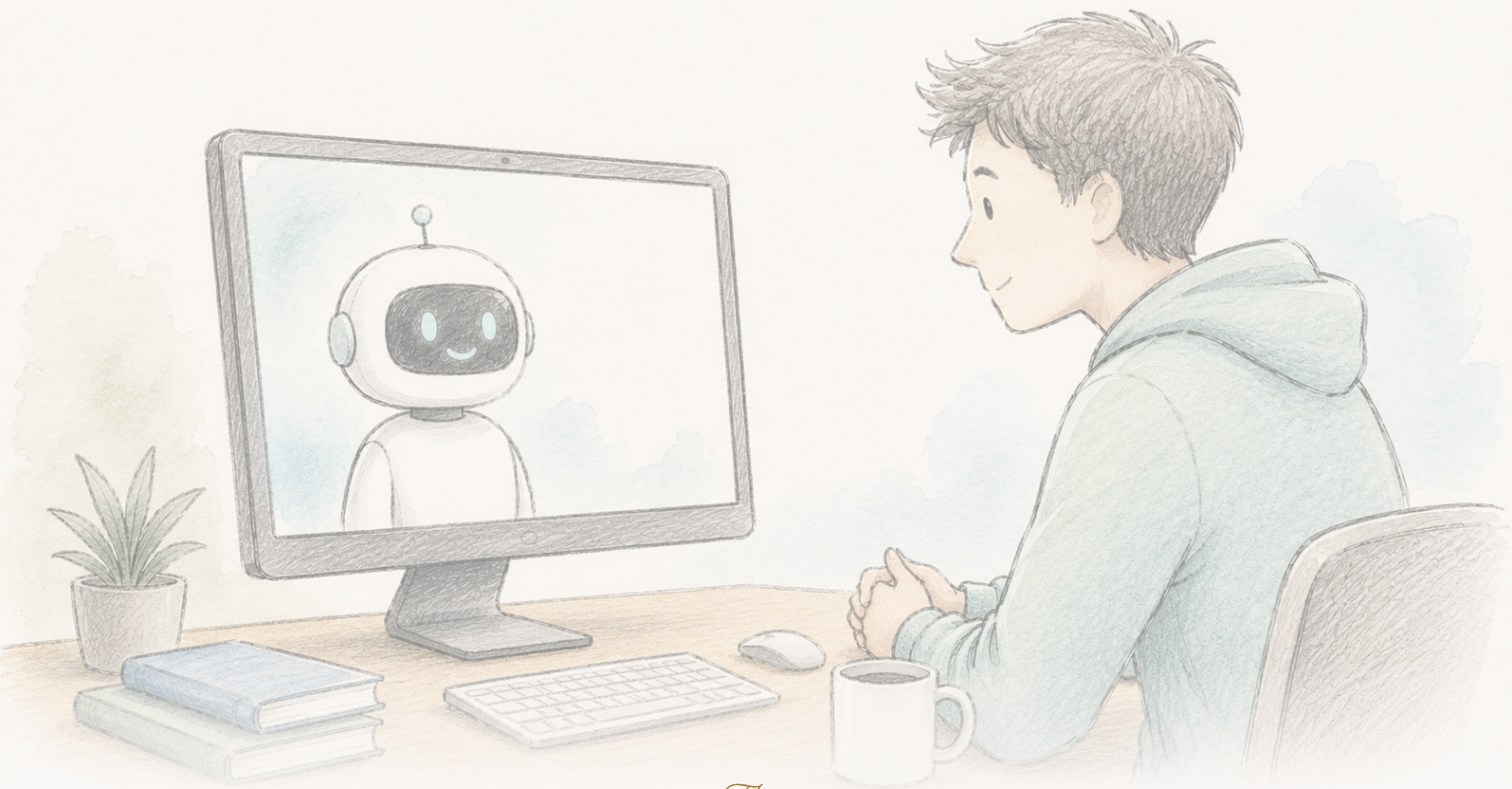




The CLEAN PAPER

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

Ένα chatbot φάνηκε να μειώνει τις συνωμοσιολογικές πεποιθήσεις για μήνες

Μελέτη του 2024 βρήκε ότι μια συνομιλία τριών γύρων με GPT-4 Turbo μείωσε κατά περίπου 20% την πίστη των ανθρώπων σε μια θεωρία συνωμοσίας που ήδη πίστευαν, με αποτέλεσμα που διατηρήθηκε δύο μήνες, εμφανίστηκε σε διαφορετικούς τύπους συνωμοσιών και βασίστηκε σε ισχυρισμούς της AI που αξιολογήθηκαν σχεδόν όλοι ως ακριβείς. Τον Ιούνιο του 2026 η εργασία έλαβε Editorial Expression of Concern για προβλήματα χειρισμού δεδομένων και αναπαραγωγιμότητας· οι συγγραφείς λένε ότι διορθωμένη ανάλυση διατηρεί την κατεύθυνση, τη στατιστική σημαντικότητα και το μέγεθος του ευρήματος, ενώ το Science συνεχίζει να αξιολογεί. Ένα πραγματικά ενδιαφέρον, προσεκτικά σχεδιασμένο αποτέλεσμα — προς το παρόν υπό εξέταση, ούτε οριστικά επιβεβαιωμένο ούτε καταρριφθέν.

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

Το Science έχει προσαρτήσει Editorial Expression of Concern στην εργασία ενώ αξιολογεί ζητήματα χειρισμού δεδομένων και αναπαραγωγιμότητας. Δεν πρόκειται για ανάκληση ούτε για διαπίστωση παραπτώματος· σημαίνει ότι το αναφερόμενο αποτέλεσμα πρέπει να διαβάζεται ως προσωρινό μέχρι να ολοκληρωθεί η εξέταση.

Editorial Expression of Concern →

Τελικά, τα γεγονότα — και μια επίσημη προειδοποίηση πάνω στην εργασία

Το 2024, μια μελέτη ανέφερε κάτι που πήγαινε κόντρα σε μια βολική κοινή πεποίθηση. Δώστε σε κάποιον μια σύντομη συνομιλία τριών γύρων με AI — GPT-4 Turbo, με οδηγία να απαντά στα συγκεκριμένα στοιχεία που ο ίδιος επικαλείται για μια θεωρία συνωμοσίας την οποία πιστεύει — και, κατά μέσο όρο, η πίστη του σε αυτή τη θεωρία πέφτει περίπου 20%. Η πτώση υπήρχε ακόμη δύο μήνες αργότερα. Το αποτέλεσμα εμφανίστηκε τόσο σε κλασικές συνωμοσίες (η δολοφονία του John F. Kennedy, εξωγήινοι, Illuminati) όσο και σε επίκαιρες (COVID-19, οι αμερικανικές εκλογές του 2020), ακόμη και σε ανθρώπους των οποίων η πεποίθηση ήταν ισχυρή και συνδεδεμένη με την ταυτότητά τους. Όταν επαγγελματίας fact-checker βαθμολόγησε 128 ισχυρισμούς που έκανε η AI, οι 127 ήταν αληθείς, ο ένας παραπλανητικός και κανένας ψευδής.

Ο τίτλος που ταξίδεψε ήταν ότι *μπορείς να βγάλεις κάποιον από το «λαγούμι» με συζήτηση* — ότι οι άνθρωποι που πιστεύουν συνωμοσίες δεν είναι απρόσιτοι στα στοιχεία, απλώς χρειάζονται τα σωστά στοιχεία, αρκετά συγκεκριμένα και προσαρμοσμένα σε αυτά που οι ίδιοι πιστεύουν. Είναι πραγματικά ενδιαφέρον ισχυρισμός και η μελέτη σχεδιάστηκε πιο προσεκτικά από το μεγαλύτερο μέρος της έρευνας για την πειθώ.

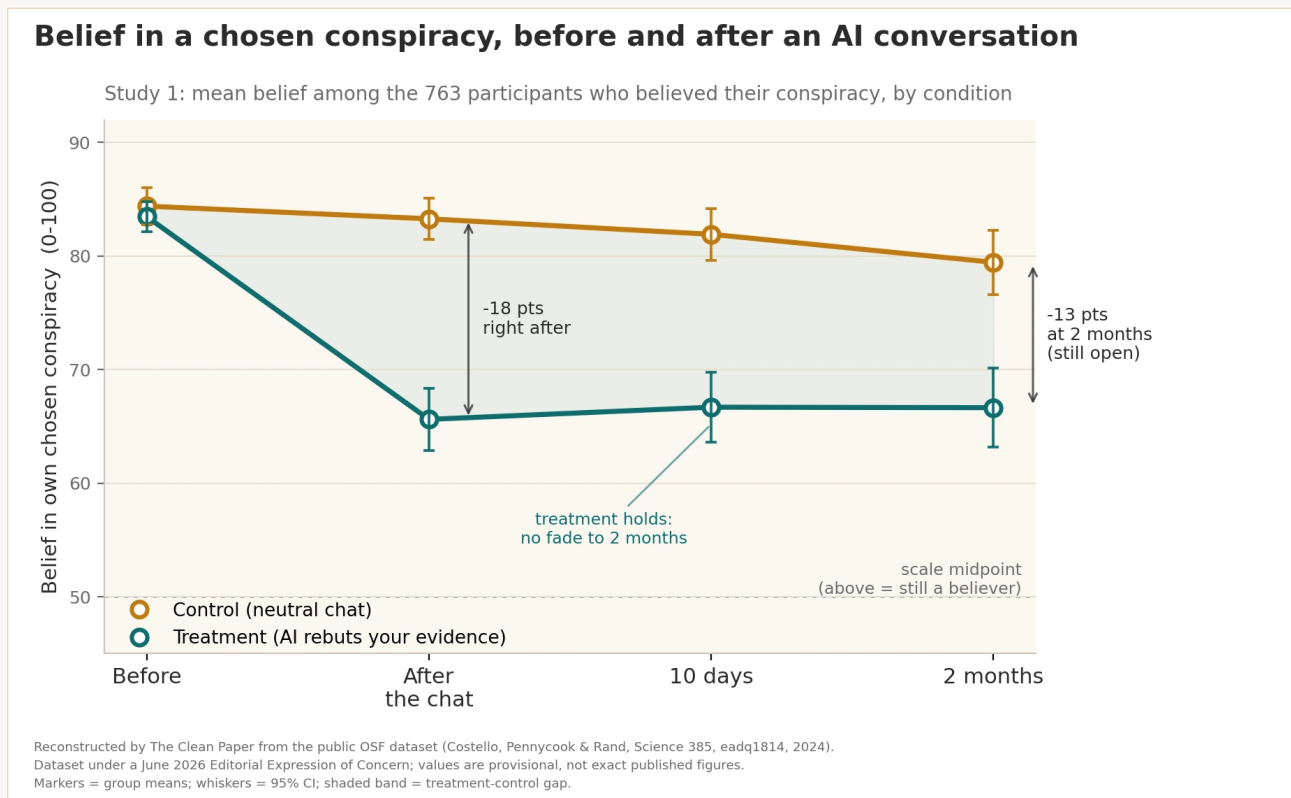
Ύστερα, τον Ιούνιο του 2026, το Science δημοσίευσε μια **Editorial Expression of Concern** για την εργασία. Αυτό είναι το σημείο που οι περισσότερες αναδηγήσεις θα παραλείψουν — και ακριβώς το σημείο που καθορίζει πόσο βάρος μπορεί να σηκώσει σήμερα το αποτέλεσμα.

Τι είναι — και τι δεν είναι — μια Expression of Concern

Μια Editorial Expression of Concern είναι επίσημη προειδοποιητική επισήμανση που ένα περιοδικό προσθέτει σε μια εργασία για να ενημερώσει τους αναγνώστες ότι έχουν τεθεί ερωτήματα τα οποία ακόμη διερευνώνται. Δεν είναι ανάκληση (η εργασία δεν έχει αποσυρθεί) και δεν αποτελεί από μόνη της διαπίστωση επιστημονικού παραπτώματος. Σημαίνει: διαβάστε το αποτέλεσμα έχοντας υπόψη αυτή την επιφύλαξη, γιατί το επιστημονικό αρχείο μπορεί να αλλάξει. Εδώ, αφού ενημερώθηκαν για τα προβλήματα, οι συγγραφείς τα διερεύνησαν, ανέφεραν τις λεπτομέρειες στο Science και παρέδωσαν διορθωμένη ανάλυση.

Αυτό που επισημάνθηκε είναι συγκεκριμένο. Οι συγγραφείς ενημερώθηκαν για ασυνέπειες στον τρόπο με τον οποίο εφαρμόζονταν τα κριτήρια διαλογής συμμετεχόντων ανάμεσα στο χειρόγραφο και στον δημοσιευμένο κώδικα ανάλυσης, καθώς και για άσχετες γραμμές που είχαν παρεισφρήσει στο δημόσιο dataset από σφάλμα συγχώνευσης κώδικα — προβλήματα που έκαναν ορισμένους από τους δημοσιευμένους αριθμούς δύσκολο να αναπαραχθούν από το υλικό που είχε δοθεί στη δημοσιότητα. Οι συγγραφείς παρέδωσαν στο *Science* διορθωμένο pipeline ανάλυσης και ενημερωμένα αποτελέσματα, τα οποία λένε ότι συμφωνούν με τα αρχικά ως προς την **κατεύθυνση, τη στατιστική σημαντικότητα και το ουσιαστικό μέγεθος**. Το *Science* τα αξιολογεί. Μέχρι να τελειώσει αυτή η αξιολόγηση, οι ακριβείς αριθμοί παραμένουν κάτω από ένα ερωτηματικό που μόνο το περιοδικό μπορεί να αφαιρέσει.

Άρα εδώ υπάρχουν δύο ιστορίες ταυτόχρονα: ένα ενδιαφέρον αποτέλεσμα για το αν τα γεγονότα μπορούν να μετακινήσουν παγιωμένες πεποιθήσεις και ένα ζωντανό παράδειγμα του επιστημονικού αρχείου που διορθώνει τον εαυτό του δημόσια. Σκοπός αυτού του άρθρου είναι να μην τις μπερδέψει.



Στη μελέτη, μια συνομιλία τριών γύρων με AI που αντέκρουε τα στοιχεία που ο ίδιος ο συμμετέχων δήλωνε ότι υποστήριζαν τη θεωρία του αναφέρθηκε ότι μείωσε κατά περίπου 20% κατά μέσο όρο την πίστη στη συγκεκριμένη συνωμοσία· σε αντίθεση με μια συνομιλία ελέγχου για άσχετο θέμα, η πτώση παρέμενε μετρήσιμη στις 10 ημέρες και στους 2 μήνες. Το αναφερόμενο αποτέλεσμα ήταν ανθεκτικό αλλά μερικό: οι περισσότεροι συμμετέχοντες εξακολουθούσαν να πιστεύουν τη θεωρία μετά — λιγότερο, όχι καθόλου. Η εργασία βρίσκεται σήμερα υπό Editorial Expression of Concern (*Science*, Ιούνιος 2026) λόγω ασυνεπειών στην αναφορά και σφάλματος στο δημόσιο dataset· οι συγγραφείς λένε ότι διορθωμένη ανάλυση διατηρεί την κατεύθυνση, τη σημαντικότητα και το μέγεθος του αποτελέσματος, ενώ το *Science* συνεχίζει την αξιολόγηση. Το διάγραμμα ανακατασκευάστηκε από το The Clean Paper με βάση αυτό το δημόσιο dataset, άρα οι τιμές είναι προσωρινές και όχι οι τελικοί αριθμοί της εργασίας.

Τι έκαναν οι συγγραφείς

- Πραγματοποίησαν δύο πειράματα με 2190 συμμετέχοντες, καθένας από τους οποίους ονόμασε — με δικά του λόγια — μια θεωρία συνωμοσίας που πίστευε και τα στοιχεία που θεωρούσε ότι την υποστήριζαν.
- Έβαλαν κάθε συμμετέχοντα να κάνει συνομιλία τριών γύρων με GPT-4 Turbo. Στην **ομάδα παρέμβασης** η ΑΙ είχε οδηγία να αντικρούει τα συγκεκριμένα στοιχεία του συμμετέχοντα· στην **ομάδα ελέγχου** συζητούσε ένα άσχετο θέμα.
- Μέτρησαν την πίστη στην επιλεγμένη συνωμοσία πριν και αμέσως μετά τη συνομιλία και επικοινωνήσαν ξανά με τους συμμετέχοντες στις **10 ημέρες και στους 2 μήνες**.
- Έβαλαν επαγγελματία fact-checker να αξιολογήσει την ακρίβεια και των 128 πραγματολογικών ισχυρισμών που έκανε η ΑΙ σε ένα δείγμα.
- Έλεγξαν αν το αποτέλεσμα επεκτεινόταν σε άλλες, άσχετες συνωμοσίες — και, ως τεστ ειδικότητας, αν μείωνε επίσης την πίστη σε συνωμοσίες που τυχαίνει να είναι αληθινές.

Τι βρήκαν

- **Μέση μείωση περίπου 20%** στην πίστη για την επιλεγμένη συνωμοσία στην ομάδα παρέμβασης σε σχέση με τον έλεγχο (μελέτη 1: διάστημα εμπιστοσύνης 95% [13.8, 19.7] μονάδες σε κλίμακα 100 βαθμών, $P < 0.001$).
- **Διατηρήθηκε.** Στην ομάδα παρέμβασης, η πτώση δεν εξαφανίστηκε: η πίστη παρέμεινε κοντά στο επίπεδο αμέσως μετά τη συνομιλία στις 10 ημέρες και στους δύο μήνες, αντί να επιστρέφει προς τα πάνω, ενώ η ομάδα ελέγχου παρέμεινε υψηλότερα σε όλη την περίοδο — η εργασία περιγράφει το αποτέλεσμα στην επιλεγμένη συνωμοσία ως ουσιαστικά αμείωτο για τουλάχιστον δύο μήνες, όχι ως στιγμιαία ταλάντωση.
- **Γενικεύτηκε** στο εύρος των συνωμοσιών που ανέφεραν οι συμμετέχοντες και ίσχυε ακόμη και όταν η αρχική πεποίθηση ήταν ισχυρή.
- **Οι ισχυρισμοί της ΑΙ ήταν ακριβείς** σε αυτό το πλαίσιο: 127 από 128 (99.2%) αξιολογήθηκαν ως αληθείς, 1 (0.8%) ως παραπλανητικός, κανένας ως ψευδής.
- **Ήταν ειδικό,** όχι γενικευμένη αμφιβολία: η παρέμβαση **δεν** μείωσε την πίστη σε αληθινές συνωμοσίες, υποδηλώνοντας ότι μετακίνησε μη υποστηριζόμενες πεποιθήσεις αντί να κάνει τους ανθρώπους κυνικούς για τα πάντα.
- **Υπήρξε μέτρια μεταφορά** — η πίστη σε άλλες, άσχετες συνωμοσίες μειώθηκε περίπου 12% και οι συμμετέχοντες δήλωσαν μεγαλύτερη πρόθεση να αντιδρούν σε συνωμοσιολογικούς ισχυρισμούς. Το κύριο αποτέλεσμα εμφανίστηκε και στη μελέτη 2 και έδειξε προς την ίδια κατεύθυνση σε ένα μικρότερο δείγμα χωρίς ομάδα ελέγχου (ασθενέστερο στοιχείο, αλλά συνεπές).

Τι δεν αποδεικνύει

- **Δεν** στέκεται σήμερα χωρίς αμφισβήτηση. Η εργασία βρίσκεται υπό Editorial Expression of Concern για προβλήματα χειρισμού δεδομένων και αναπαραγωγιμότητας, και οι διορθωμένοι αριθμοί ακόμη αξιολογούνται από το Science. Αντιμετωπίστε τις συγκεκριμένες τιμές ως προσωρινές μέχρι να ολοκληρωθεί η εξέταση.
- **Δεν** δείχνει ότι η ΑΙ «θεραπεύει» τη συνωμοσιολογική πίστη. Μια μέση μείωση ~20% είναι ουσιαστική μετατόπιση, όχι εξάλειψη — οι περισσότεροι συμμετέχοντες εξακολουθούσαν να πιστεύουν τη θεωρία μετά, απλώς λιγότερο.
- **Δεν** είναι αποτέλεσμα πραγματικής ανάπτυξης στον κόσμο. Οι συμμετέχοντες επέλεξαν να συμμετάσχουν σε μελέτη και συνεργάστηκαν καλόπιστα με την ΑΙ· έξω από το εργαστήριο, όσοι είναι πιο δεσμευμένοι σε μια συνωμοσία είναι συχνά και οι λιγότερο πιθανό να αναζητήσουν chatbot που θα αντιπαρατεθεί μαζί τους.
- Η ακρίβεια 99.2% είναι **ειδική για αυτή την εργασία**, αυτό το **μοντέλο και αυτή την προσεκτική καθοδήγηση** — όχι γενική εγγύηση ότι τα γλωσσικά μοντέλα λένε αλήθεια. Οι συγγραφείς τονίζουν ότι η ίδια εξατομικευμένη δύναμη πειθούς θα μπορούσε να ενισχύσει ψευδείς πεποιθήσεις αν στοχευόταν προς αυτή την κατεύθυνση.
- «Ανθεκτικό» εδώ σημαίνει **δύο μήνες**, κάτι εντυπωσιακό για μία συνομιλία, αλλά όχι το ίδιο με μόνιμο.

Πόσο ισχυρά είναι τα στοιχεία

- **Ο σχεδιασμός είναι πραγματικό πλεονέκτημα.** Ελεγχόμενα πειράματα παρέμβασης έναντι ελέγχου, καθαρός έλεγχος τύπου placebo, επανάληψη σε δύο μελέτες και ανεξάρτητο δείγμα, follow-up διάρκειας και ρητός έλεγχος της ακρίβειας των ίδιων των ισχυρισμών της ΑΙ — είναι πιο προσεκτική δουλειά από μεγάλο μέρος της έρευνας για την πειθώ, και η κατεύθυνση του αποτελέσματος είναι συνεπής όπου ελέγχθηκε.
- **Αλλά η Expression of Concern δεν είναι υποσημείωση.** Το να μπορεί κανείς να ξαναπαράγει τους δημοσιευμένους αριθμούς από τα διαθέσιμα δεδομένα και τον κώδικα είναι μέρος αυτού που κάνει ένα αποτέλεσμα αξιόπιστο, και αυτό ακριβώς παρουσίασε πρόβλημα — ασυνέπειες στα κριτήρια διαλογής και διπλές/άσχετες γραμμές από σφάλμα συγχώνευσης κώδικα. Η δήλωση των συγγραφέων ότι το διορθωμένο pipeline διατηρεί το αποτέλεσμα είναι ενθαρρυντική και εύλογη, αλλά παραμένει η δική τους αναφορά, όχι ακόμη ανεξάρτητη ετυμηγορία. Η αξιολόγηση του Science είναι αυτή που μετρά.
- **Η έντιμη κατάσταση είναι «επισημασμένο», όχι «καταρρίφθηκε» ούτε «επιβεβαιώθηκε».** Μια προσεκτικά σχεδιασμένη, εντυπωσιακή μελέτη της οποίας οι ακριβείς αριθμοί βρίσκονται υπό επίσημη εξέταση. Είναι άβολη θέση για να αφήσει κανείς μια καλή ιστορία, αλλά είναι η ακριβής.

Γιατί έχει σημασία

Για χρόνια, μια επιδραστική άποψη υποστήριζε ότι όσοι πιστεύουν θεωρίες συνωμοσίας κινούνται από ψυχολογικές ανάγκες και ταυτότητα και επομένως δεν μπορούν να μετακινηθούν με επιχειρήματα και γεγονότα. Μια ανθεκτική, βασισμένη σε στοιχεία μείωση — αν επιβεβαιωθεί — είναι ουσιαστικό αντίβαρο σε αυτή την απαισιοδοξία: υποδηλώνει ότι μέρος του προβλήματος ήταν πως η γενική αποδόμηση δεν ασχολούνταν ποτέ με τα συγκεκριμένα στοιχεία που επικαλείται ένας πιστός, κάτι που ένα LLM μπορεί να κάνει σε μεγάλη κλίμακα.

Έχει επίσης σημασία ως παράδειγμα του πώς υποτίθεται ότι λειτουργεί η επιστήμη. Το μάθημα μιας Expression of Concern δεν είναι ότι τα διάσημα αποτελέσματα δεν αξίζουν τίποτα· είναι ότι τα λάθη αναδεικνύονται, τα δεδομένα ξαναεξετάζονται και οι ισχυρισμοί επανελέγχονται δημόσια. Το ότι αυτός ο μηχανισμός λειτουργεί είναι χαρακτηριστικό, όχι σκάνδαλο — και είναι ο λόγος που η σωστή στάση απέναντι στο αποτέλεσμα είναι υπομονή, όχι ούτε χειροκρότημα ούτε απόρριψη.

Και για την ΑΙ το νόμισμα έχει δύο όψεις. Η ίδια ικανότητα που μπορεί να αποδυναμώσει μια ψευδή πεποίθηση με εξατομικευμένο επιχείρημα θα μπορούσε εξίσου αποτελεσματικά να ενισχύσει μια άλλη. Η τεχνική είναι ουδέτερη· ο σκοπός όχι.

Καθαρή σύνοψη

Μελέτη του 2024 βρήκε ότι μια συνομιλία τριών γύρων με GPT-4 Turbo μείωσε κατά περίπου 20% την πίστη των ανθρώπων σε μια θεωρία συνωμοσίας που ήδη πίστευαν, αποτέλεσμα που διατηρήθηκε δύο μήνες και γενικεύτηκε σε διαφορετικούς τύπους συνωμοσιών, ενώ οι πραγματολογικοί ισχυρισμοί της ΑΙ αξιολογήθηκαν σχεδόν όλοι ως ακριβείς. Τον Ιούνιο του 2026 η εργασία τέθηκε υπό Editorial Expression of Concern για προβλήματα χειρισμού δεδομένων και αναπαραγωγιμότητας· οι συγγραφείς αναφέρουν ότι διορθωμένη ανάλυση διατηρεί την κατεύθυνση, τη σημαντικότητα και το μέγεθος του ευρήματος, και το *Science* συνεχίζει να αξιολογεί. Διαβάστε το ως ένα πραγματικά ενδιαφέρον, προσεκτικά χτισμένο αποτέλεσμα που σήμερα βρίσκεται υπό εξέταση — όχι οριστικά επιβεβαιωμένο, όχι καταρριφθέν. Η πιο έντιμη πρόταση για αυτό είναι αυτή που γράφει η ίδια η διαδικασία: ελέγξτε το ξανά.

Πηγές

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>