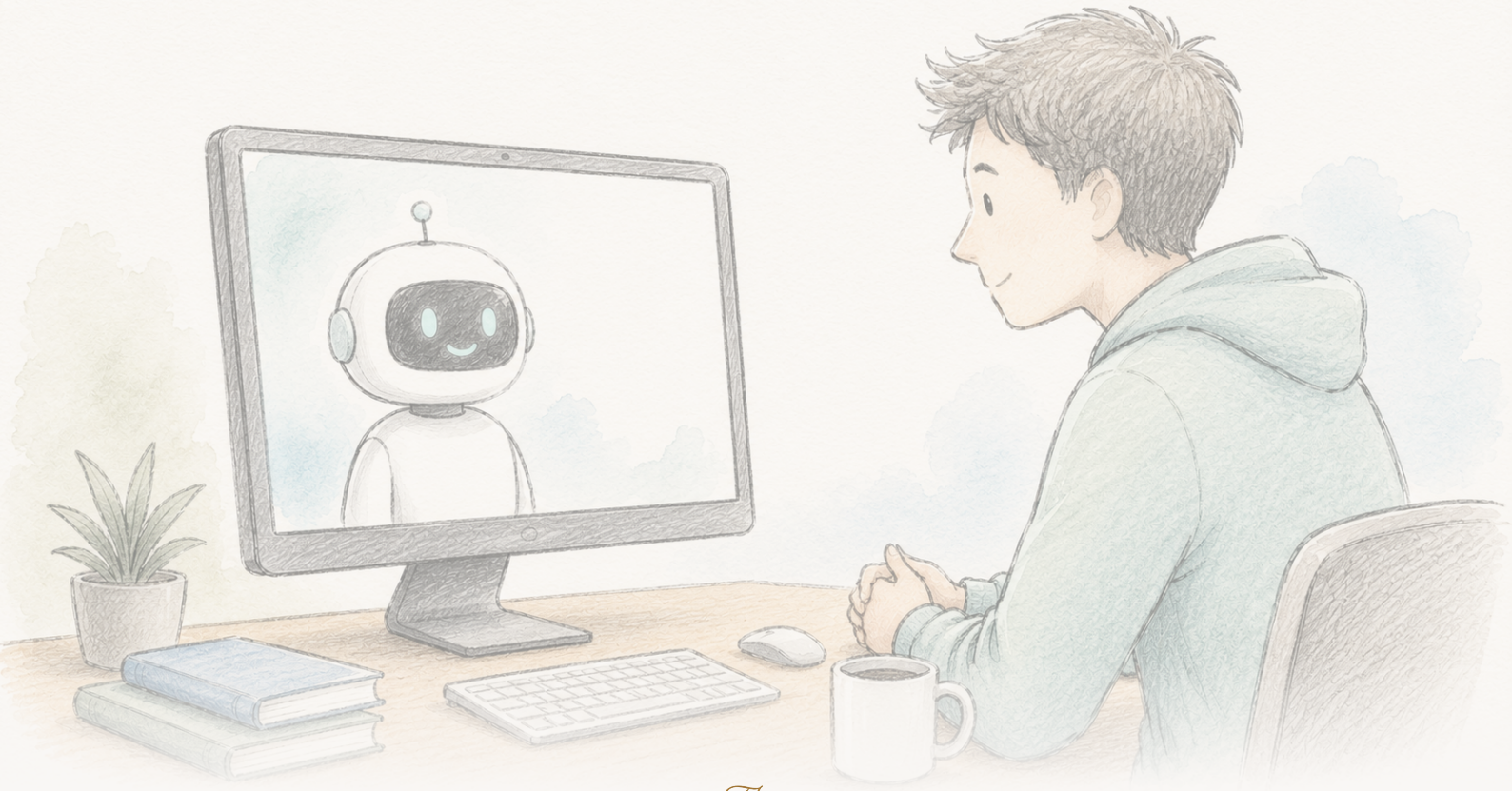


The Clean Paper

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

ჩატბოტმა კონსპირაციული რწმენა, როგორც ჩანს, რამდენიმე თვით შეამცირა

2024 წლის კვლევაში GPT-4 Turbo-სთან პერსონალიზებულმა, მტკიცებულებაზე დაფუძნებულმა საუბარმა კონსპირაციული რწმენა საშუალოდ დაახლოებით 20%-ით შეამცირა, და ეფექტი ორი თვის შემდეგაც ჩანდა. 2026 წლის ივნისში კი Science-მა მონაცემების დამუშავებისა და გამეორებადობის საკითხების გამო სტატიას Editorial Expression of Concern დაურთო. ავტორები ამბობენ, რომ შესწორებული ანალიზი ძირითად შედეგს ინარჩუნებს; ჟურნალი მას ჯერ კიდევ აფასებს. შედეგი ამჟამად განხილვაშია — არც საბოლოოდ დადასტურებული და არც უარყოფილი.

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-მა სტატიას Editorial Expression of Concern დაურთო მონაწილეთა სკრინინგის კრიტერიუმებთან და საჯარო მონაცემთა ნაკრებში კოდის გაერთიანების შეცდომით მოხვედრილ ზედმეტ სტრიქონებთან დაკავშირებული შეუსაბამობების გამო. ეს არ არის რეტრაქცია. ავტორებმა შესწორებული ანალიზი წარადგინეს და ამბობენ, რომ ძირითადი შედეგის მიმართულება, სტატისტიკური მნიშვნელოვნება და ზომა შენარჩუნებულია; Science-ის შეფასება გრძელდება, ამიტომ კონკრეტული რიცხვები დროებითია.

Editorial Expression of Concern →

ფაქტები მაინც მოქმედებს — და სტატიაზე გაფრთხილების ნიშანი

2024 წელს გამოქვეყნებულმა კვლევამ საკმაოდ კომფორტულ გავრცელებულ შეხედულებას შეუპირისპირა შედეგი. ადამიანს თუ მისცემთ AI-სთან მოკლე, სამრაუნდიან საუბარს — GPT-4 Turbo-ს, რომელსაც მითითებული აქვს უპასუხოს სწორედ იმ მტკიცებულებებს, რომლებსაც მონაწილე პირადად სჯერა თავისი კონსპირაციული თეორიის დასადასტურებლად — საშუალოდ ამ თეორიისადმი რწმენა დაახლოებით 20%-ით მცირდება. ეს შემცირება ორი თვის შემდეგაც შენარჩუნდა. ეფექტი გამოჩნდა როგორც კლასიკურ კონსპირაციებზე (ჯონ ფ. კენედის მკვლელობა, უცხოპლანეტელები, ილუმინატები), ისე აქტუალურ თემებზე (COVID-19, აშშ-ის 2020 წლის არჩევნები) და მათ შორის იმ ადამიანებშიც, რომელთა რწმენაც ძლიერი და იდენტობასთან დაკავშირებული იყო. პროფესიონალმა ფაქტების შემმოწმებელმა AI-ის მიერ გაკეთებული 128 ფაქტობრივი მტკიცება შეაფასა: 127 იყო ჭეშმარიტი, ერთი — შეცდომაში შემყვანი, ხოლო არცერთი — მცდარი.

გავრცელებული სათაური ასეთი იყო: ადამიანთან ფაქტებზე საუბარს მაინც შეუძლია მისი „კურდღლის სოროდან“ გამოყვანა — კონსპირაციული რწმენის მქონე ადამიანები მტკიცებულებებისთვის მიუწვდომლები არ არიან; უბრალოდ საჭიროა ზუსტად ის მტკიცებულება, რომელიც მათ საკუთარ არგუმენტს პასუხობს. ეს ნამდვილად საინტერესო დასკვნაა და კვლევაც დარწმუნების თემაზე შესრულებული ბევრი ნაშრომისგან განსხვავებით საკმაოდ ფრთხილად იყო აგებული.

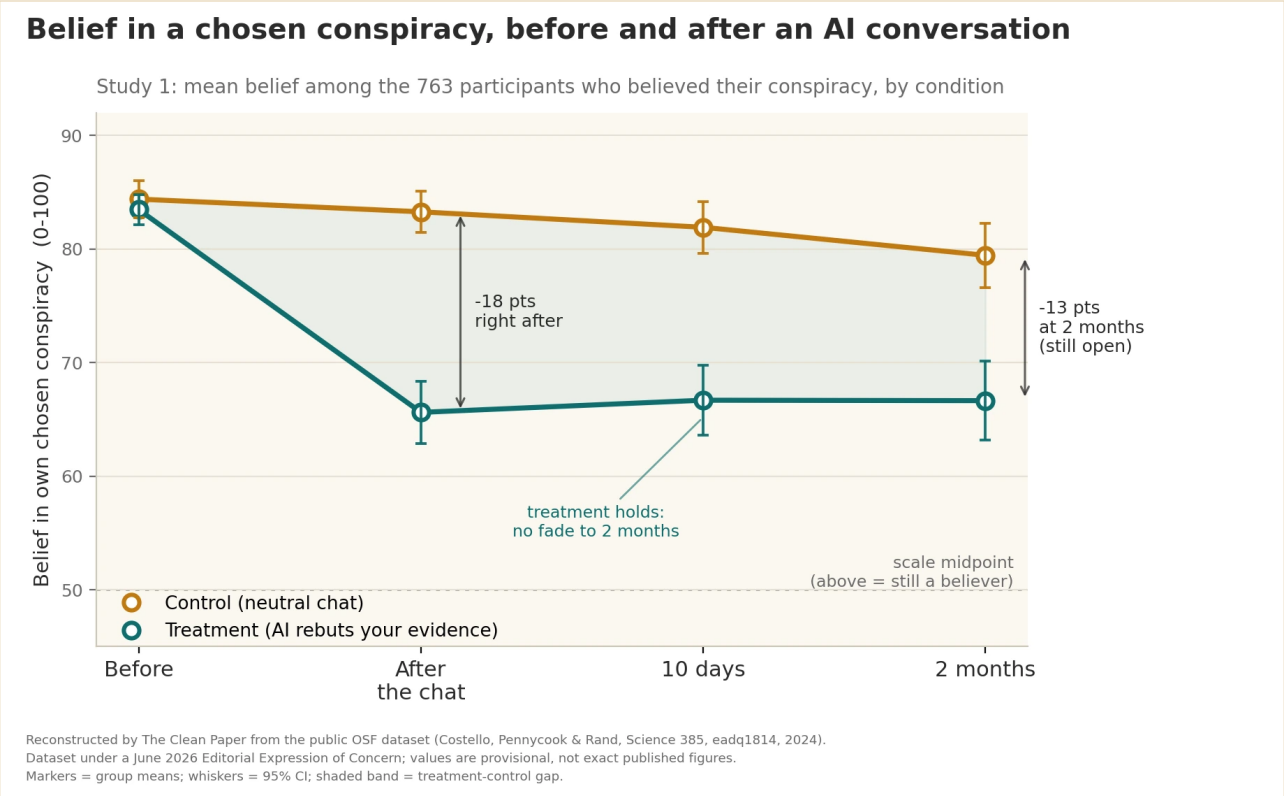
შემდეგ, 2026 წლის ივნისში, *Science*-მა სტატიას **Editorial Expression of Concern** — სარედაქციო შეშფოთების განცხადება — დაურთო. სწორედ ეს ნაწილია, რომელსაც ბევრი გადმოცემა გამოტოვებს, თუმცა ამჟამად შედეგისთვის დასაშვები ნდობის დონეს სწორედ ის განსაზღვრავს.

რას ნიშნავს Expression of Concern — და რას არ ნიშნავს

Editorial Expression of Concern არის ოფიციალური გაფრთხილება, რომელსაც ჟურნალი სტატიას ურთავს, როცა მის შესახებ კითხვები გაჩნდა და მათი გარკვევა ჯერ მიმდინარეობს. ეს **არ არის** რეტრაქცია (სტატია უკან არ არის გაწვეული) და თავისთავად არც არაკეთილსინდისიერების დადგენას ნიშნავს. მისი აზრია: ეს ნაშრომი წაიკითხეთ ამ შენიშვნის გათვალისწინებით, რადგან სამეცნიერო ჩანაწერი შეიძლება შეიცვალოს. ამ შემთხვევაში ავტორებმა პრობლემების შესახებ ინფორმაციის მიღების შემდეგ საკითხი შეისწავლეს, დეტალები *Science*-ს აცნობეს და შესწორებული ანალიზი წარადგინეს.

გაფრთხილება კონკრეტულ საკითხებს ეხება. ავტორებს აცნობეს, რომ ხელნაწერსა და გამოქვეყნებულ ანალიზის კოდში მონაწილეთა სკრინინგის კრიტერიუმები თანმიმდევრულად არ იყო გამოყენებული; გარდა ამისა, საჯარო მონაცემთა ნაკრებში კოდის გაერთიანების შეცდომის შედეგად ზედმეტი სტრიქონები მოხვდა. ამის გამო გამოქვეყნებული მასალებიდან ზოგიერთი რიცხვის ზუსტად გამეორება რთული გახდა. ავტორებმა *Science*-ს გადასცეს შესწორებული ანალიზის პროცესი და განახლებული შედეგები, რომლებიც, მათი თქმით, საწყის შედეგებს **მიმართულებით, სტატისტიკური მნიშვნელოვნებით და არსებითი ეფექტის ზომით** ემთხვევა. *Science* მათ ახლა აფასებს. ამ შეფასების დასრულებამდე კონკრეტული რიცხვები პირობითად უნდა მივიჩნიოთ — საბოლოო პასუხის გაცემა მხოლოდ ჟურნალს შეუძლია.

ამიტომ აქ ერთდროულად ორი ამბავია: საინტერესო შედეგი იმის შესახებ, შეუძლია თუ არა ფაქტებს გამყარებული რწმენის შეცვლა, და ცოცხალი მაგალითი იმისა, როგორ ასწორებს სამეცნიერო ჩანაწერი საკუთარ თავს საჯაროდ. ამ სტატიის მიზანია, ეს ორი ამბავი ერთმანეთში არ აგვერიოს.



კვლევაში მონაწილესთან სამრაუნდიანი AI-საუბარი, რომელიც მის მიერ დასახელებულ მტკიცებულებებს პასუხობდა, საშუალოდ დაახლოებით 20%-ით ამცირებდა მის მიერ არჩეული კონსპირაციული თეორიისადმი რწმენას. დაუკავშირებელ თემაზე საკონტროლო საუბრისგან განსხვავებით, შემცირება 10 დღის და 2 თვის შემდეგაც იზომებოდა. ეფექტი ხანგრძლივი, მაგრამ არასრული იყო: მონაწილეთა უმეტესობას კონსპირაციის კვლავ სჯეროდა — რწმენა შემცირდა, არ გამქრალა. სტატიას ამჟამად *Science*-ის სარედაქციო „Expression of Concern“ (2026 წლის ივნისი) აქვს დართული ანგარიშგების შეუსაბამობებისა და საჯარო მონაცემთა ნაკრებში შეცდომის გამო. ავტორები ამბობენ, რომ შესწორებული ანალიზი ეფექტის მიმართულებას, სტატისტიკურ მნიშვნელოვნებასა და ზომას ინარჩუნებს, ხოლო *Science* შეფასებას აგრძელებს. ეს დიაგრამა The Clean Paper-მა საჯარო მონაცემთა ნაკრებიდან აღადგინა, ამიტომ ნაჩვენები მნიშვნელობები დროებითია და არა სტატიის საბოლოო რიცხვები.

Original figure — The Clean Paper CC BY 4.0

რა გააკეთეს ავტორებმა

- ჩაატარეს ორი ექსპერიმენტი 2190 მონაწილით. თითოეულმა საკუთარი სიტყვებით დაასახელა კონსპირაციული თეორია, რომლისაც სჯეროდა, და მტკიცებულება, რომელიც, მისი აზრით, ამ რწმენას ამყარებდა.
- თითოეულ მონაწილეს GPT-4 Turbo-სთან სამრაუნდიანი საუბარი ჰქონდა. **სამკურნალო/ინტ**
  პირობებში AI-ს ევალებოდა მონაწილეს მიერ დასახელებული კონკრეტული მტკიცებულებების გაბათილება; **საკონტროლო** პირობებში კი საუბარი დაუკავშირებელ თემაზე მიმდინარეობდა.
- არჩეული კონსპირაციისადმი რწმენა გაზომეს საუბრამდე და დაუყოვნებლივ მის

შემდეგ, შემდეგ კი მონაწილეებს **10 დღის და 2 თვის** შემდეგ კვლავ დაუკავშირდნენ.
- პროფესიონალ ფაქტების შემმოწმებელს შეაფასებინეს შერჩეულ ნიმუშში AI-ის მიერ გაკეთებული 128-ვე ფაქტობრივი მტკიცების სიზუსტე.
- შეამოწმეს, ვრცელდებოდა თუ არა ეფექტი სხვა, დაუკავშირებელ კონსპირაციებზე და — სპეციფიკურობის ტესტად — ამცირებდა თუ არა ის იმ კონსპირაციებისადმი რწმენასაც, რომლებიც რეალურად **ჭეშმარიტია**.

რა აღმოაჩინეს

- ინტერვენციის ჯგუფში არჩეული კონსპირაციისადმი რწმენა საკონტროლო ჯგუფთან შედარებით **საშუალოდ დაახლოებით 20%-ით შემცირდა** (კვლევა 1: 100-ქულიან სკალაზე 95%-იანი სანდოობის ინტერვალი [13.8, 19.7] ქულა, $P < 0.001$).
- **ეფექტი შენარჩუნდა.** ინტერვენციის ჯგუფში შემცირება არ გაქრა: 10 დღის და ორი თვის შემდეგ რწმენა თითქმის იმავე დონეზე დარჩა, რაც საუბრის დასრულებისთანავე იყო, მაშინ როცა საკონტროლო ჯგუფში მთელი პერიოდის განმავლობაში უფრო მაღალი მაჩვენებელი იყო. ავტორები არჩეულ კონსპირაციაზე ეფექტს აღწერენ როგორც მინიმუმ ორი თვის განმავლობაში უცვლელად შენარჩუნებულს და არა წამიერ რყევას.
- **ეფექტი სხვადასხვა კონსპირაციაზე ვრცელდებოდა** და შენარჩუნდა იმ მონაწილეებშიც, რომელთა საწყისი რწმენა ძლიერი იყო.
- **ამ კონკრეტულ პირობებში AI-ის მტკიცებები ზუსტი იყო:** 128-დან 127 (99.2%) შეფასდა ჭეშმარიტად, 1 (0.8%) — შეცდომაში შემყვანად, არცერთი — მცდარად.
- **ეფექტი სპეციფიკური იყო და არა ზოგადი უნდობლობა:** ინტერვენციამ ჭეშმარიტი კონსპირაციებისადმი რწმენა **არ** შეამცირა. ეს მიუთითებს, რომ ის დაუსაბუთებელ რწმენებს შეეხო და არა ყველაფრის მიმართ ცინიზმს.
- **იყო ზომიერი გადადებითი ეფექტიც** — სხვა, დაუკავშირებელი კონსპირაციებისადმი რწმენა დაახლოებით 12%-ით შემცირდა და მონაწილეები ამბობდნენ, რომ მომავალში კონსპირაციულ მტკიცებებს უფრო მეტად შეეწინააღმდეგებოდნენ. ძირითადი ეფექტი მეორე კვლევაშიც გამეორდა და იმავე მიმართულებას აჩვენებდა უფრო პატარა ნიმუშშიც, რომელსაც საკონტროლო ჯგუფი არ ჰქონდა (ეს უფრო სუსტი, მაგრამ თავსებადი მტკიცებულებაა).

რას არ ამტკიცებს ეს კვლევა

- ამჟამად შედეგი **ეჭვის გარეშე მიღებული არ არის**. სტატიას მონაცემების დამუშავებისა და გამეორებადობის პრობლემების გამო Editorial Expression of Concern აქვს დართული, ხოლო შესწორებულ რიცხვებს *Science* ჯერ კიდევ აფასებს. ამ მიმოხილვის დასრულებამდე კონკრეტული მნიშვნელობები დროებითად უნდა ჩაითვალოს.
- კვლევა **არ** აჩვენებს, რომ AI კონსპირაციულ რწმენას „კურნავს“. საშუალოდ ~20%-იანი შემცირება მნიშვნელოვანი ცვლილებაა, მაგრამ არა გაქრობა — მონაწილეთა

უმეტესობას თეორიის კვლავ სჯეროდა, უბრალოდ ნაკლებად.
- ეს **არ არის რეალურ სამყაროში დანერგვის შედეგი**. მონაწილეები კვლევაში ნებაყოფლობით ჩაერთნენ და AI-ს კეთილსინდისიერად ესაუბრებოდნენ; რეალურ ცხოვრებაში ყველაზე მტკიცე რწმენის მქონე ადამიანები შეიძლება ყველაზე ნაკლებად ეძებდნენ ჩატბოტს, რომელიც მათ შეეწინააღმდეგება.
- 99.2%-იანი სიზუსტე **ამ კონკრეტულ ამოცანას, მოდელსა და ფრთხილ პრომპტინგს ეხება** — ეს არ არის ზოგადი გარანტია, რომ ენობრივი მოდელები ყოველთვის სიმართლეს ამბობენ. თავად ავტორები ხაზს უსვამენ, რომ იგივე პერსონალიზებული დამარწმუნებელი ძალა შეიძლება მცდარი რწმენის გასაძლიერებლადაც გამოიყენონ.
- „ხანგრძლივი“ აქ **ორ თვეს** ნიშნავს. ერთ საუბართან შედარებით ეს შთამბეჭდავია, მაგრამ მუდმივ ეფექტს არ უდრის.

რამდენად ძლიერია მტკიცებულება

- **კვლევის დიზაინი ნამდვილად ძლიერი მხარეა.** კონტროლირებული ინტერვენცია-საკონტროლო ექსპერიმენტები, კარგი პლაცებოს მსგავსი კონტროლი, ორი კვლევისა და დამოუკიდებელი ნიმუშის გამეორება, ხანგრძლივობის შემოწმება და AI-ის საკუთარი მტკიცებების სიზუსტის ცალკე შეფასება — ეს დარწმუნების კვლევების უმეტესობაზე უფრო ფრთხილი დიზაინია, ხოლო ეფექტის მიმართულება ყველა ტესტში ერთნაირი იყო.
- **მაგრამ Expression of Concern უბრალო სქოლიო არ არის.** გამოქვეყნებული მონაცემებიდან და კოდიდან რიცხვების ხელახლა მიღების შესაძლებლობა შედეგის სანდოობის ნაწილია, და სწორედ აქ გაჩნდა პრობლემა — სკრინინგის კრიტერიუმების შეუსაბამობა და კოდის გაერთიანების შეცდომით დუბლირებული/ზედმეტი სტრიქონები. ავტორების განცხადება, რომ შესწორებული პროცესი შედეგს ინარჩუნებს, დამაიმედებელი და შესაძლებლად დამაჯერებელია, მაგრამ ეს ჯერ ავტორების საკუთარი ანგარიშია და არა დამოუკიდებელი საბოლოო შეფასება. გადამწყვეტი იქნება *Science*-ის შეფასება.
- **დღევანდელი ზუსტი სტატუსია „გაფრთხილებით მონიშნული“, და არა „უარყოფილი“ ან „დადასტურებული“.** ეს არის კარგად აგებული და თვალსაჩინო კვლევა, რომლის ზუსტი რიცხვებიც ოფიციალურ განხილვაშია. კარგი ამბის ასეთ გაურკვეველ მდგომარეობაში დატოვება უსიამოვნოა, მაგრამ ამჟამად სწორედ ეს არის ზუსტი აღწერა.

რატომ არის ეს მნიშვნელოვანი

წლების განმავლობაში გავლენიანი შეხედულება ამბობდა, რომ კონსპირაციული რწმენა ფსიქოლოგიური საჭიროებებითა და იდენტობით არის განპირობებული და ამიტომ ადამიანს მხოლოდ ფაქტებით ამ რწმენიდან ვერ გამოიყვანს. თუ ეს შედეგი საბოლოო შემოწმებას გაუძლებს, ხანგრძლივი, მტკიცებულებაზე დაფუძნებული შემცირება ამ პესიმიზმის მნიშვნელოვანი საპირწონე იქნება: შეიძლება პრობლემა

ნაწილობრივ ის იყო, რომ ზოგადი „ფაქტ-ჩეკი“ არასოდეს პასუხობდა **იმ კონკრეტულ მტკიცებულებას**, რომელსაც კონკრეტული ადამიანი ეყრდნობა — LLM-ს კი ამის გაკეთება მასშტაბურად შეუძლია.

ეს კვლევა მნიშვნელოვანია იმის საჩვენებლადაც, თუ როგორ უნდა მუშაობდეს მეცნიერება. Expression of Concern-ის გაკვეთილი ის კი არ არის, რომ ცნობილი შედეგები არაფრად ღირს; არამედ ის, რომ შეცდომები საჯაროდ ჩნდება, მონაცემები ხელახლა მოწმდება და მტკიცებები ისევ გადის შემოწმებას. ამ მექანიზმის მუშაობა სისტემის ძლიერი მხარეა და არა სკანდალი — ამიტომ ამ შედეგთან სწორი დამოკიდებულება მოთმინებაა, არა არც დაუყოვნებელი აპლოდისმენტები და არც უარყოფა.

AI-ს მხრივაც დასკვნა ორმხრივია. იგივე უნარი, რომელიც მორგებული არგუმენტით მცდარ რწმენას ასუსტებს, შეიძლება მცდარი რწმენის გასაძლიერებლადაც ისეთივე ეფექტური იყოს. მეთოდი ნეიტრალურია; მიზანი — არა.

სუფთა შეჯამება

2024 წლის კვლევაში GPT-4 Turbo-სთან სამრაუნდიანმა საუბარმა ადამიანებში მათ მიერ არჩეული კონსპირაციული თეორიისადმი რწმენა საშუალოდ დაახლოებით 20%-ით შეამცირა; ეფექტი ორი თვის განმავლობაში შენარჩუნდა, სხვადასხვა ტიპის კონსპირაციაზე გავრცელდა და AI-ის ფაქტობრივი მტკიცებები თითქმის მთლიანად ზუსტად შეფასდა. 2026 წლის ივნისში მონაცემების დამუშავებისა და გამეორებადობის პრობლემების გამო სტატიას Editorial Expression of Concern დაურთეს. ავტორები ამბობენ, რომ შესწორებული ანალიზი შედეგის მიმართულებას, სტატისტიკურ მნიშვნელოვნებასა და ზომას ინარჩუნებს, ხოლო *Science* შეფასებას ჯერ კიდევ აგრძელებს. ამიტომ ეს უნდა წავიკითხოთ როგორც ნამდვილად საინტერესო და ფრთხილად აგებული შედეგი, რომელიც ამჟამად შემოწმების პროცესშია — არც საბოლოოდ დადასტურებული და არც უარყოფილი. ამ კვლევის შესახებ ყველაზე ზუსტ წინადადებას ახლა თავად სამეცნიერო პროცესი წერს: **ჯერ არ ვიცით საბოლოო პასუხი.**

წყაროები

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