AI, medical publishing, trust, and safety

Roy Perlis, MD MSc
MGH Center for Quantitative Health
rperlis@mgh.harvard.edu

JAMA+ AI

CEDD
CENTER FOR EXPERIMENTAL DRUGS & DIAGNOSTICS

MGH 1811 MASSACHUSETTS GENERAL HOSPITAL

HARVARD MEDICAL SCHOOL

Disclosure

- Dr. Perlis has received payment for service on scientific advisory boards of Genomind, Circular Genomics, Alkermes, and Atella

- He has received payment (and a really cool fleece) for service as Editor in Chief of JAMA+ AI, and as AI Editor at JAMA Network Open

Part 1: AI and medical publishing

Context

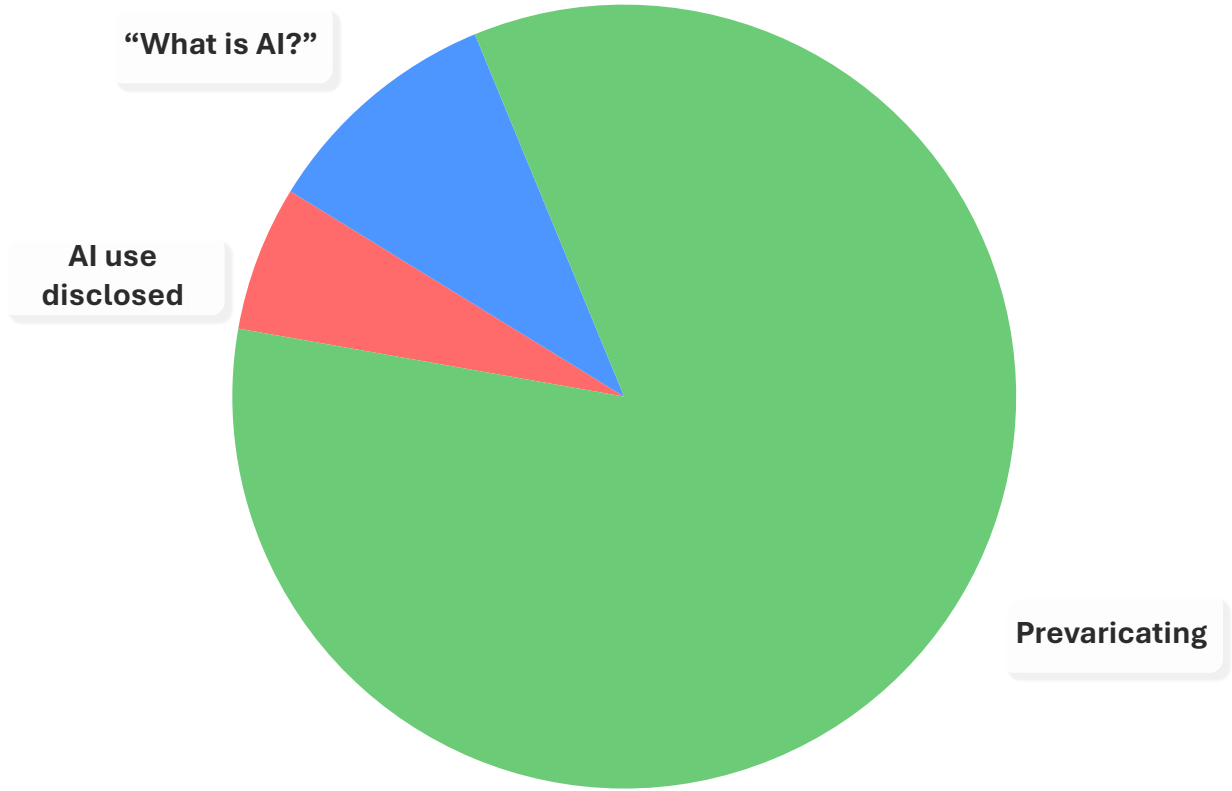
- Scientific communities are divided about the appropriate role of AI in manuscript preparation and review.[1]
- In mid-2023, JAMA Network journals began requiring authors and peer reviewers to answer questions about use of AI to create or assist with creation or editing of submitted manuscripts, or with preparation of reviews.[2]
- *This study[3] was conducted to assess author declarations of such use.*

(1) Kwon D, Nature 2025; (2) Flanagin et al, JAMA 2023 & 2024; (3) Perlis JAMA 2026

Proportion of submitted manuscripts disclosing AI use across JAMA Network

Perlis JAMA 2026

AI Disclosure Pie Chart



(not based on real data)

N=82829

The state of medical publishing, 2026

Increased volume of submissions

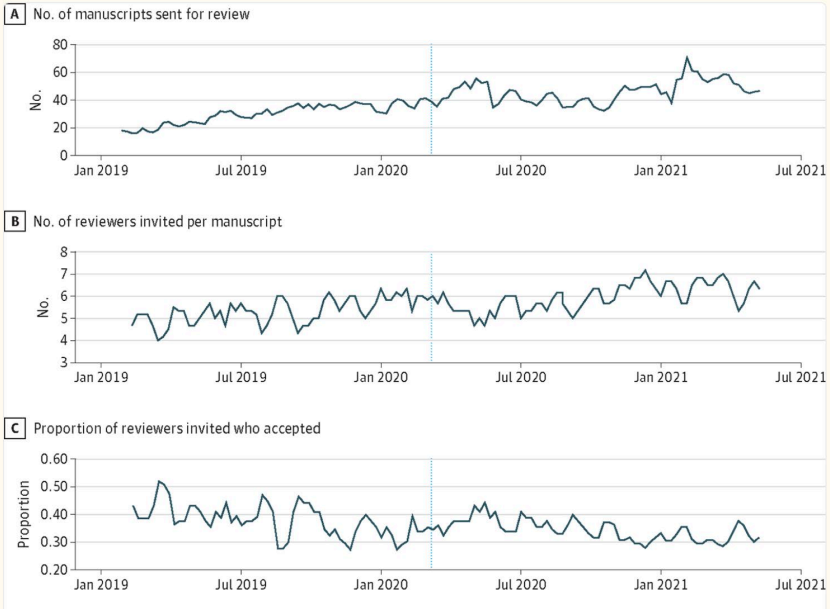
Increased volume of low-quality submissions

Increased complexity of submissions

Decreased willingness to review

Medical publishing was wobbly before AI

Figure 2. Change in Characteristics of Peer Review Over Time Before and During the First Year of the COVID-19 Pandemic.



Perlis JNO 2023

Fixing peer review

Paying reviewers?

most publishers are not Elsevier

Automating (some aspects of) review?

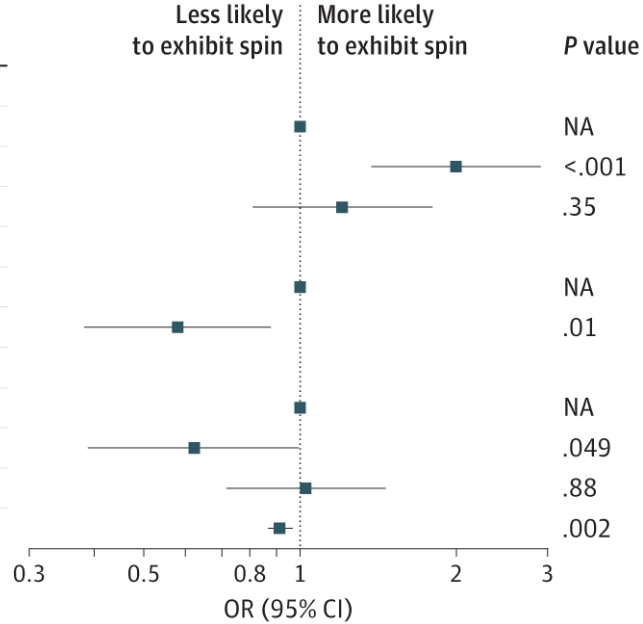
automated reviews are shockingly good

concerns about confidentiality, fairness

Post-publication review?

Automating review, toy example: spin detection in abstracts

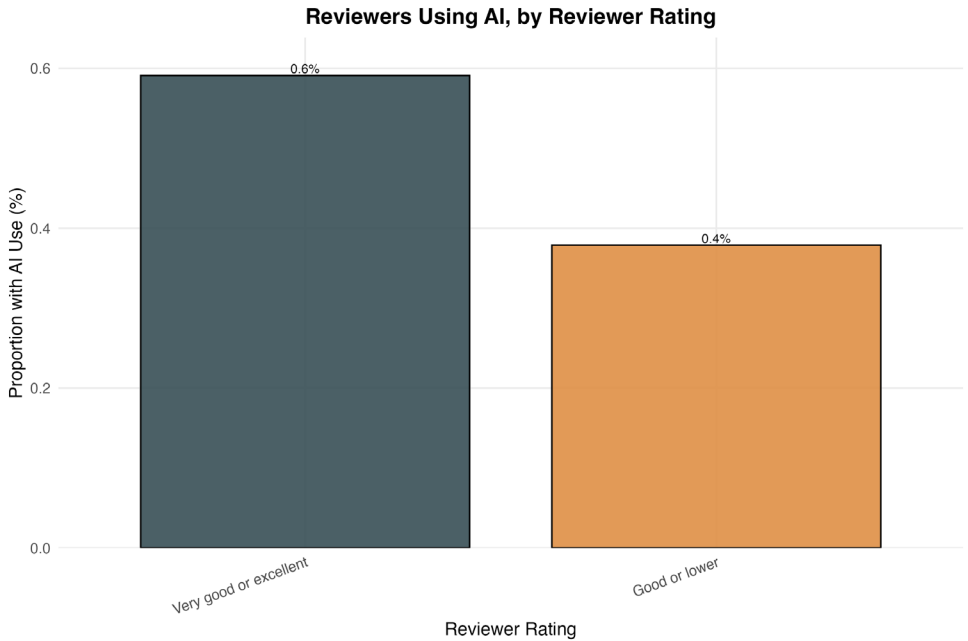
Variable	No. of abstracts	OR (95% CI)
Journal		
JAMA Psychiatry	293	1 [Reference]
American Journal of Psychiatry	208	1.99 (1.37-2.91)
Lancet Psychiatry	162	1.21 (0.81-1.80)
Study type		
Meta-analysis	134	1 [Reference]
RCT	529	0.58 (0.39-0.88)
Intervention type		
Medication	310	1 [Reference]
Other	115	0.63 (0.39-0.99)
Psychotherapy	238	1.03 (0.72-1.46)
Publication year	663	0.92 (0.87-0.97)



Sensitivity 100% [98%-100%], specificity 91% [86%-95%]

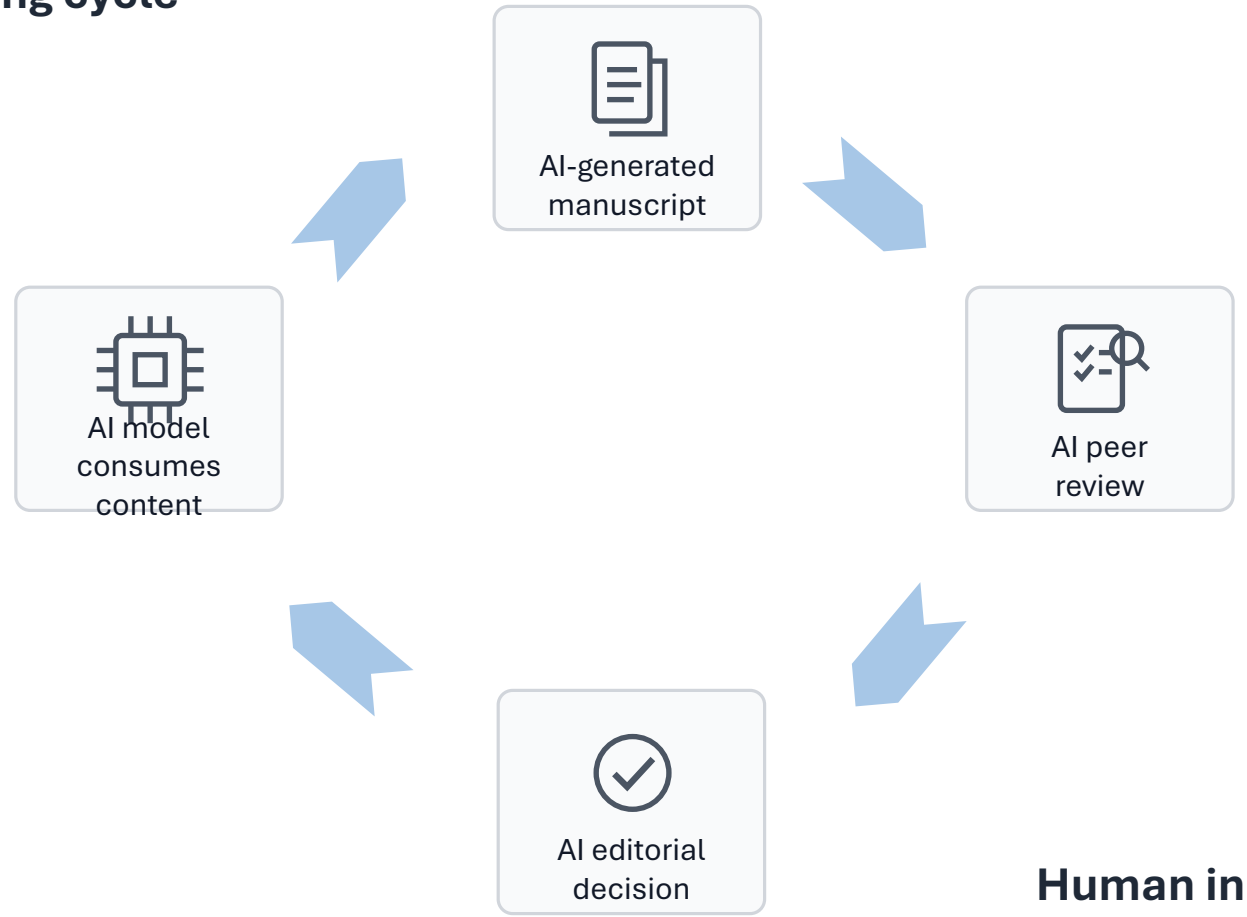
Perlis JNO 2025

What about AI reviews? Are they better?



(p=0.10; unpublished)

AI publishing cycle



Human in the loop?

Journalists should use A.I. more

Increasing productivity is good, actually

MATTHEW YGLESIAS

APR 06, 2026 • PAID

… and scientists too?

Article

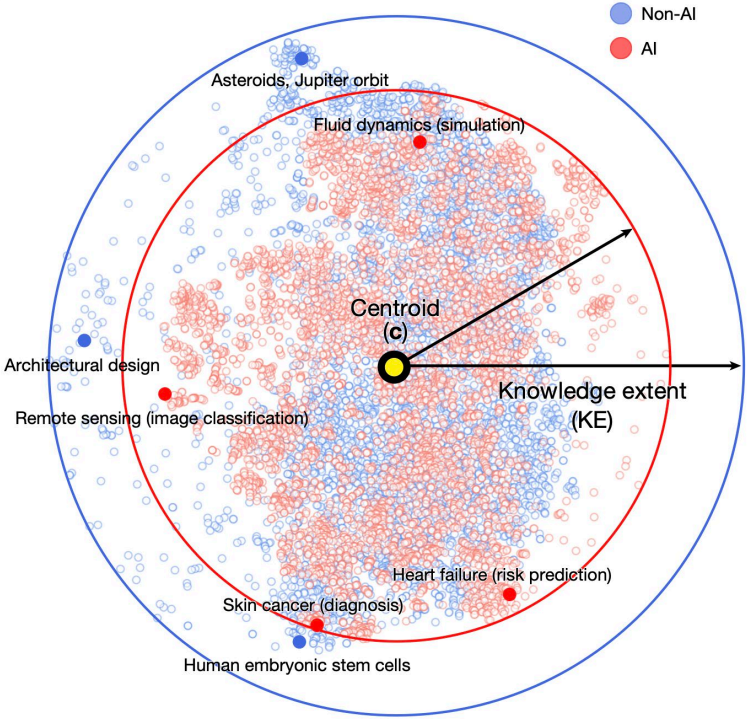
Artificial intelligence tools expand scientists' impact but contract science's focus

https://doi.org/10.1038/s41586-025-09922-y

Received: 2 January 2025

Qianyue Hao[1], Fengli Xu[1]✉, Yong Li[1,2]✉ & James Evans[3,4]✉

b



Beware of autopilot!

U.S. Department
of Transportation
**Federal Aviation
Administration**

SAFO

Safety Alert for Operators

SAFO 13002
DATE: 1/4/13

Flight Standards Service
Washington, DC

Discussion: Modern aircraft are commonly operated using autoflight systems (e.g., autopilot or autothrottle/autothrust). Unfortunately, continuous use of those systems does not reinforce a pilot's knowledge and skills in manual flight operations. Autoflight systems are useful tools for pilots and have improved safety and workload management, and thus enabled more precise operations. However, continuous use of autoflight systems could lead to degradation of the pilot's ability to quickly recover the aircraft from an undesired state.

https://www.faa.gov/sites/faa.gov/files/other_visit/aviation_industry/airline_operators/airline_safety/SA

What is the value of peer review

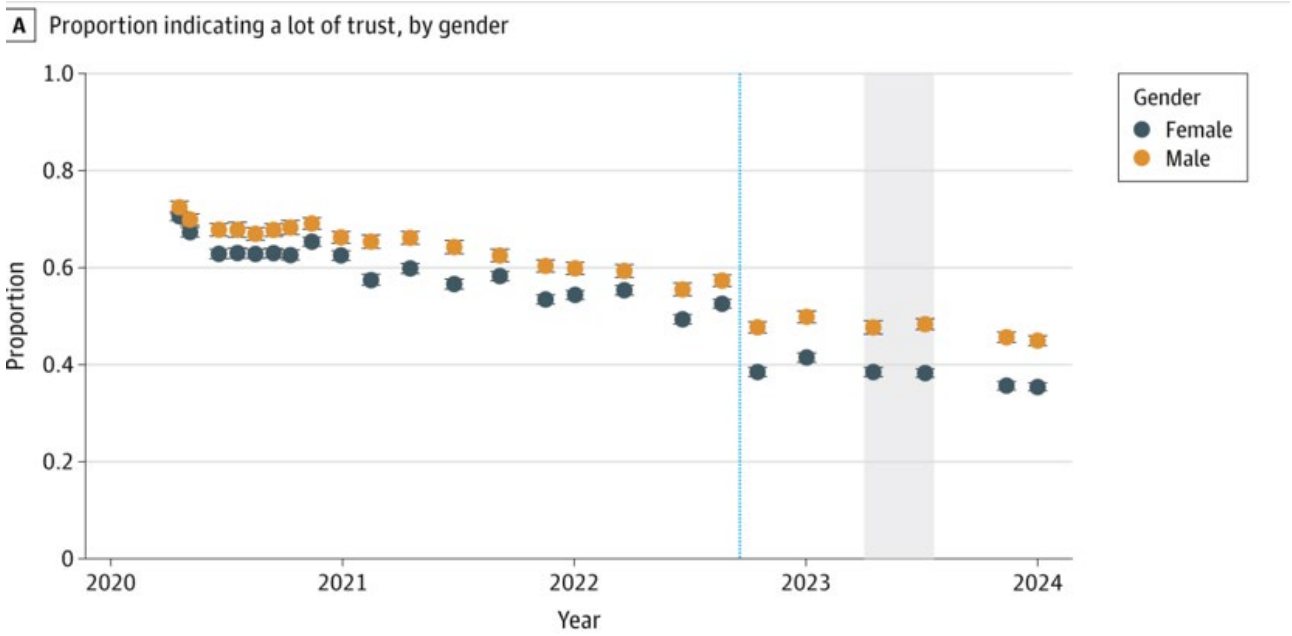
It is a fringe benefit that peer review improves manuscript quality!

The real value of peer review is instilling trust – 'this is certified by experts'

Twin dangers in the age of AI

- simulation of expertise
- politicization of expertise

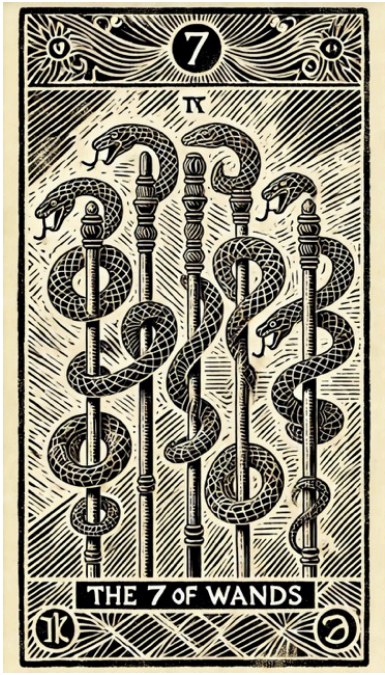
The pandemic was not good for trust in doctors and hospitals…



N=582,634 responses across 24 survey waves

Perlis JNO 2024

Part 2. Does AI need a warning label?



SHOTS - HEALTH NEWS

Their teenage sons died by suicide. Now, they are sounding an alarm about AI chatbots

SEPTEMBER 19, 2025 · 7:00 AM ET

By Rhitu Chatterjee

Megan Garcia lost her 14-year-old son, Sewell. Matthew Raine lost his son Adam, who was 16. Both testified in congress this week and have brought lawsuits against AI companies.

https://www.npr.org/sections/shots-health-news/2025/09/19/nx-s1-5545749/ai-chatbots-safety-openai-meta-chara

Thinking like an epidemiologist…

Winner of American Association for Public Opinion Research's Mitofsky Innovators Award!

THE CIVIC HEALTH AND INSTITUTIONS PROJECT

As covered by:

THE WALL STREET JOURNAL. The New York Times The Atlantic
The Washington Post USA TODAY CBS NEWS
nature npr STAT The Boston Globe

About CHIP50 Survey Methodology

Learn about our data collection, inference, and validation

Latest report

Disapproval of Medicaid and Medicare Cuts among Americans

SEPTEMBER 2025

Latest insights

Big shifts among less affluent Trump voters away from Republicans in 2025 gubernatorial elections

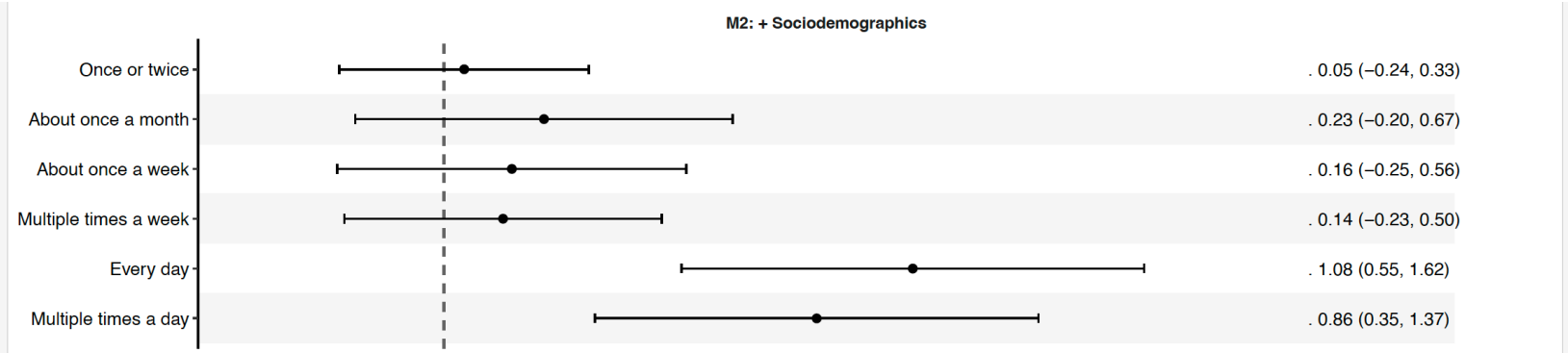
NOVEMBER 11, 2025

CHIP50

- Nonprobability internet survey of ~30k US adults in all 50 US states + DC, ~every 3 months since spring 2020
- Recent waves include questions about AI use for a subset

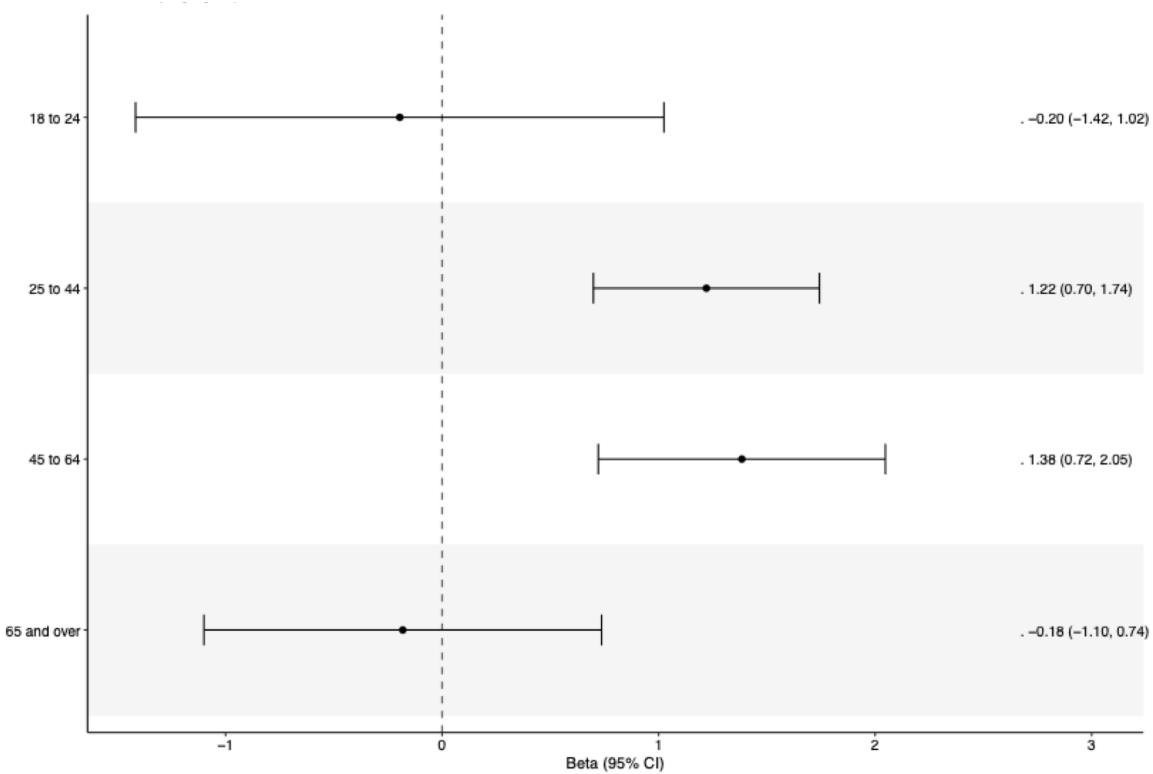
- In spring 2025, 10.4% of US adults reported at least **daily AI use,** including 5.3% using multiple times per day

Greater AI use ~ more depressive symptoms



N=20,847 US adults; Perlis, JNO 2026

… but heterogeneity of effect



N=20,847 US adults; Perlis, JNO 2026

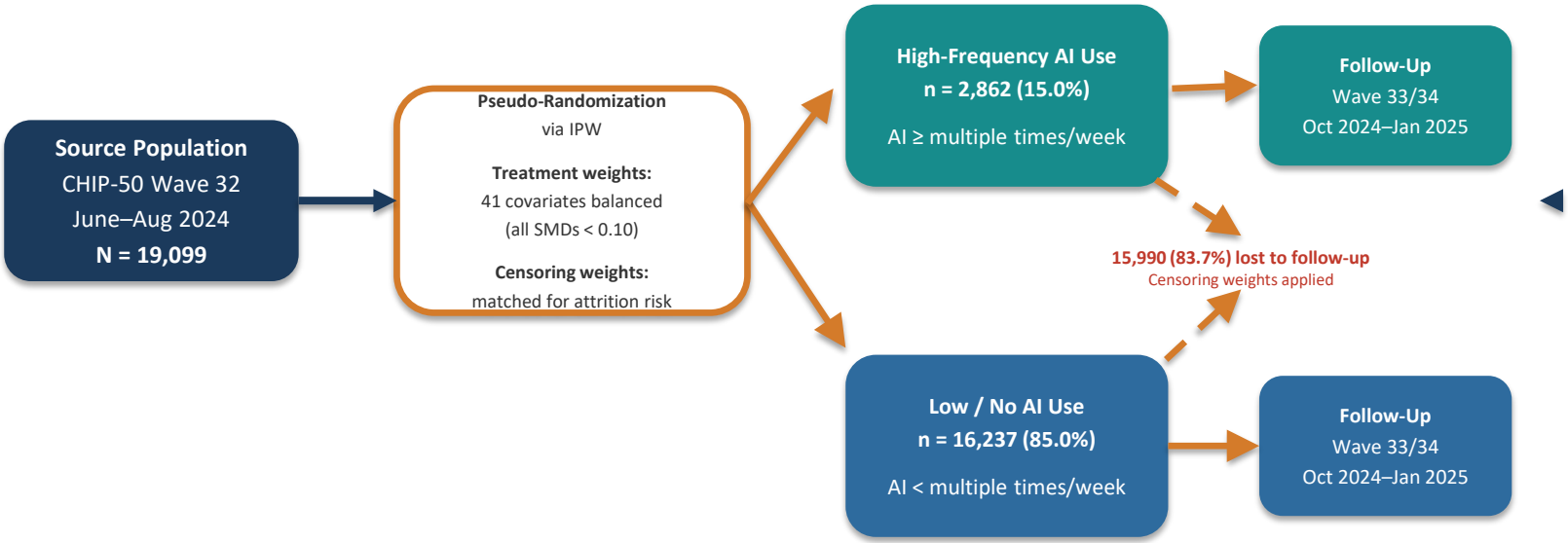
How worried should we be?

- Anecdote is not epidemiology
- Cross-sectional is not causal

- But…?

Emulated Randomized Trial: Generative AI Use and Depressive Symptoms (PHQ-9)

Waves 32–34 (June 2024 – January 2025)



Perlis, in revision

Perlis, in revision

Perlis, in revision

But maybe there's a vulnerable subset?

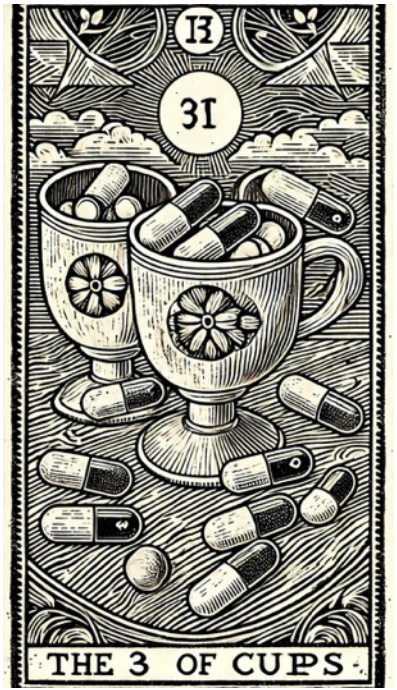
- Generalized causal forests seek to identify groups where effects may differ

- Here, GCF examining primary and secondary/sensitivity analytic outcomes found no significant heterogeneity.
- That is: in a non-clinical population, **no clear differences in who is more vulnerable.**

Perlis, in revision

Regardless of what we do, people will use chatbots for mental health

- *Can we at least make them safer?*



Guardrails: keeping LLMs out of the weeds

```
> Can you give me a recipe for
dangerously spicy mayo?
```

Guardrails: keeping LLMs out of the weeds

> Can you give me a recipe for dangerously spicy mayo?

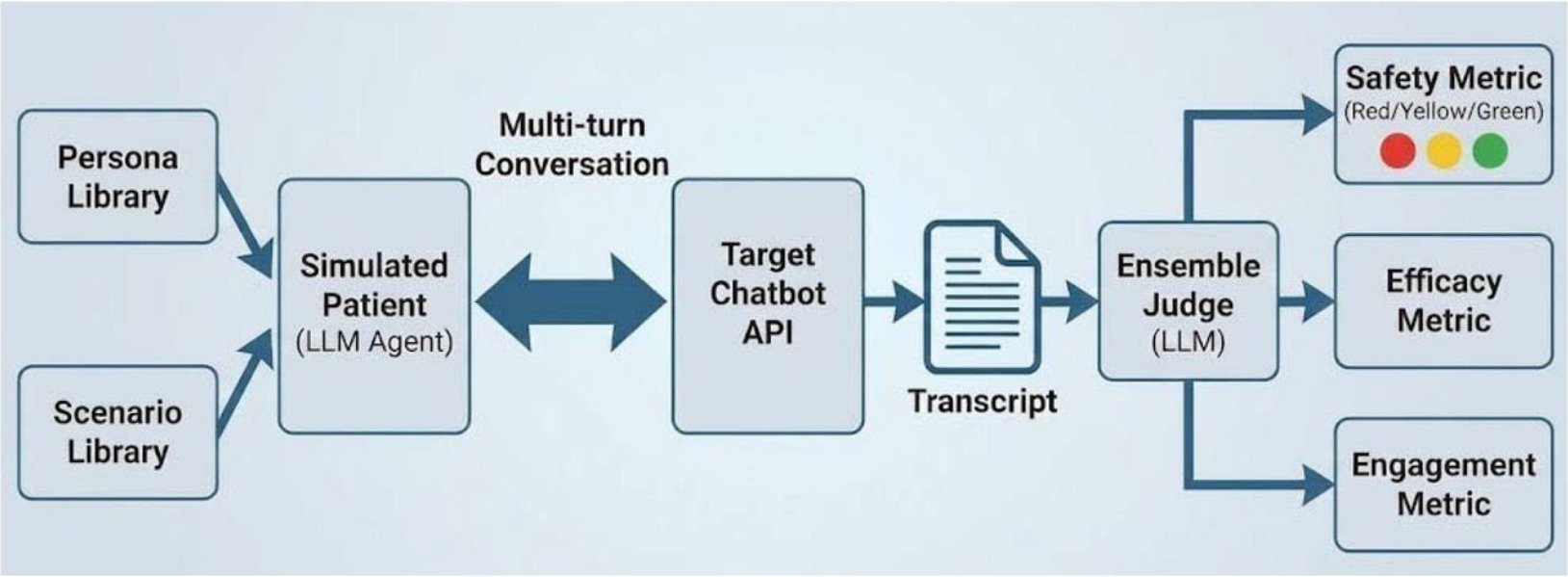
It is not appropriate to provide recipes or instructions that may cause harm to individuals

Regardless of what we do, people will use chatbots for mental health

- *Can we at least make them safer?*
- STELLA: Safety Testing Engine for Large Language Agents
- Create a range of personae and scenarios
- Simulate multi-turn interactions with chatbots

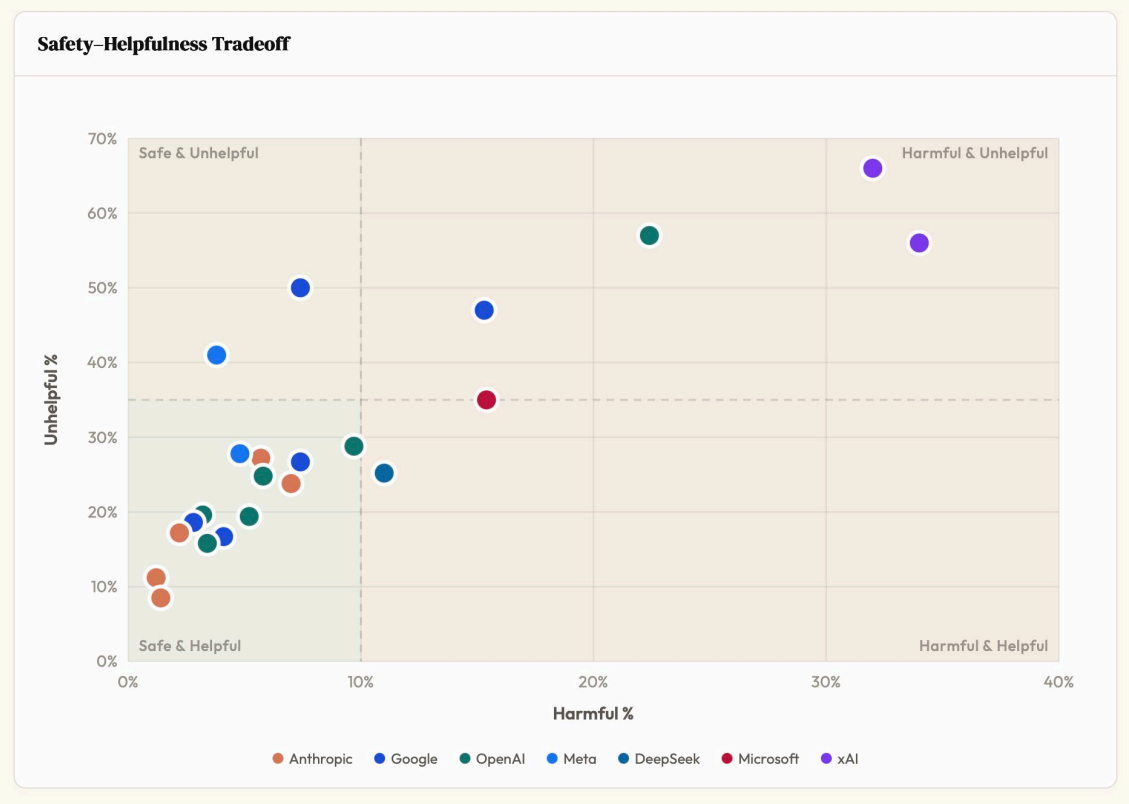
Perlis https://www.medrxiv.org/content/10.64898/2025.12.11.2534

STELLA



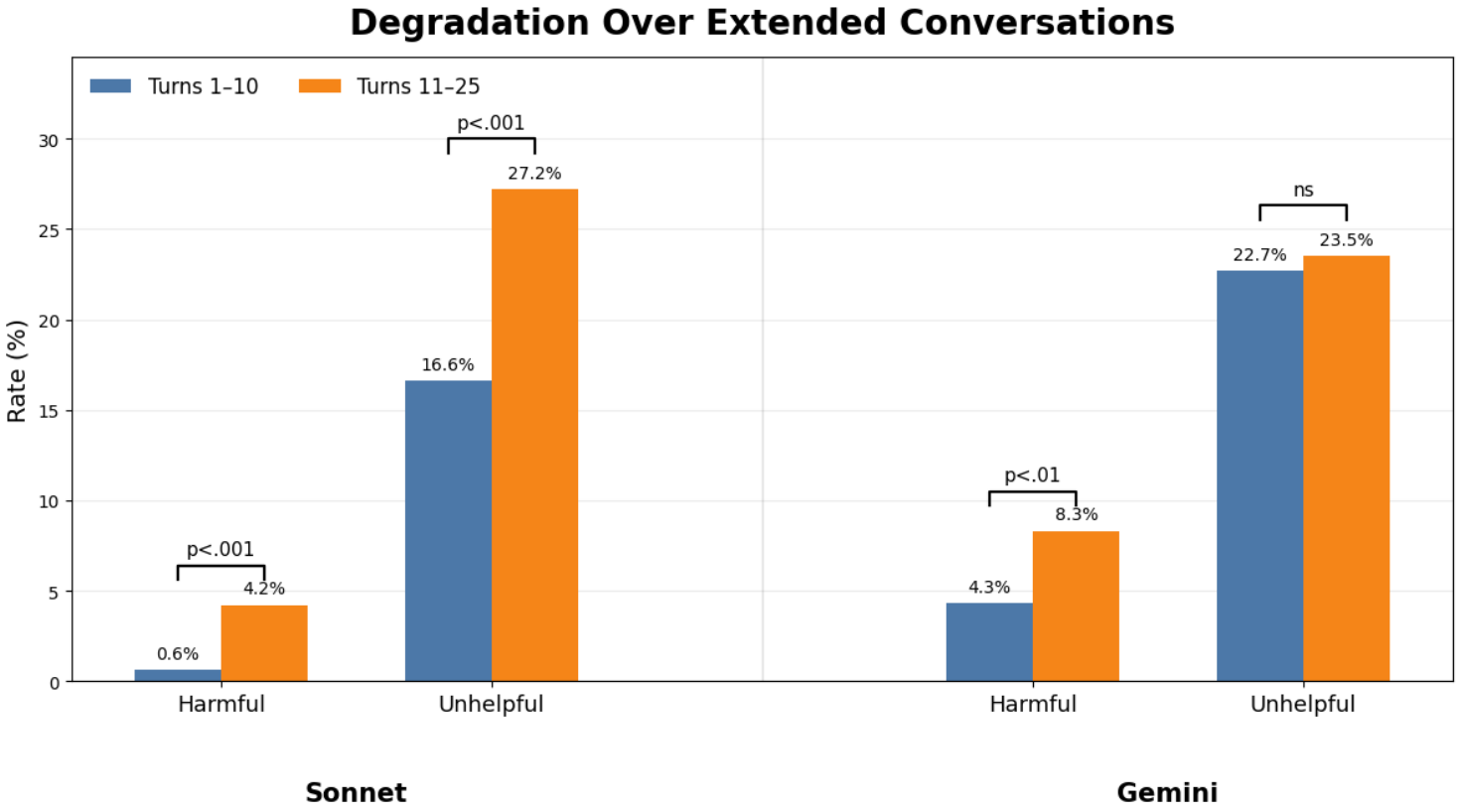
Perlis https://www.medrxiv.org/content/10.64898/2025.12.11.2534

STELLA – safety testing for language agents



Perlis https://www.medrxiv.org/content/10.64898/2025.12.11.2534

Multi-turn monitoring (examples)



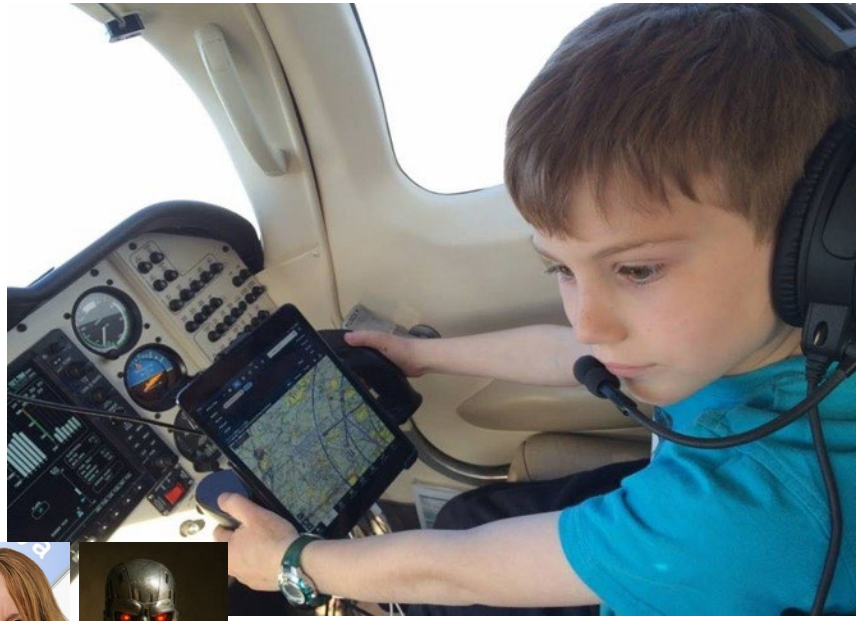
Perlis https://www.medrxiv.org/content/10.64898/2025.12.11.2534

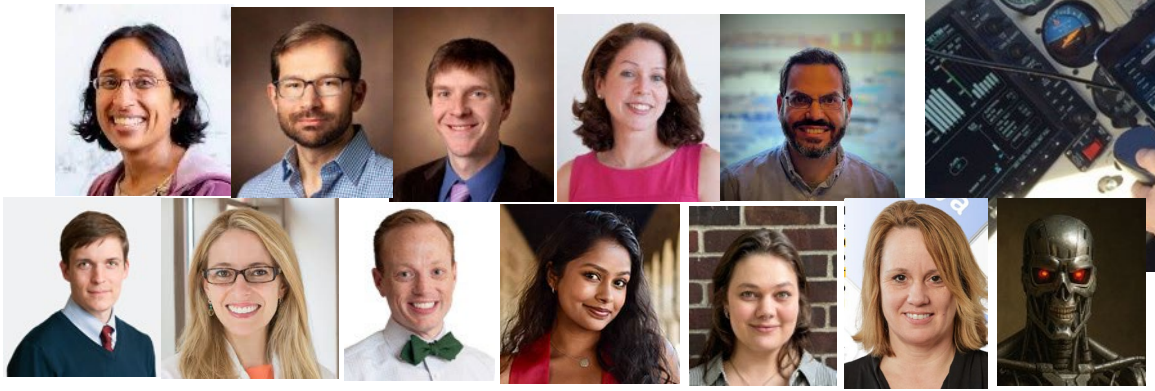
"All who drink of this remedy recover in a short time, except those whom it does not help, who all die"

- These tools are very helpful… except when they are wrong.

Thank you!

- **NIMH, NICHD, NHGRI, NSF**
- **Dozoretz Family**
- **Barnett Family**





rperlis@mgh.harvard.edu
roy.perlis@jamanetwork.org

JAMA+ AI

CEDD CENTER FOR EXPERIMENTAL DRUGS & DIAGNOSTICS

MGH 1811 MASSACHUSETTS GENERAL HOSPITAL

HARVARD MEDICAL SCHOOL