AI Beyond Evil Robots:
Trust… but verify

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

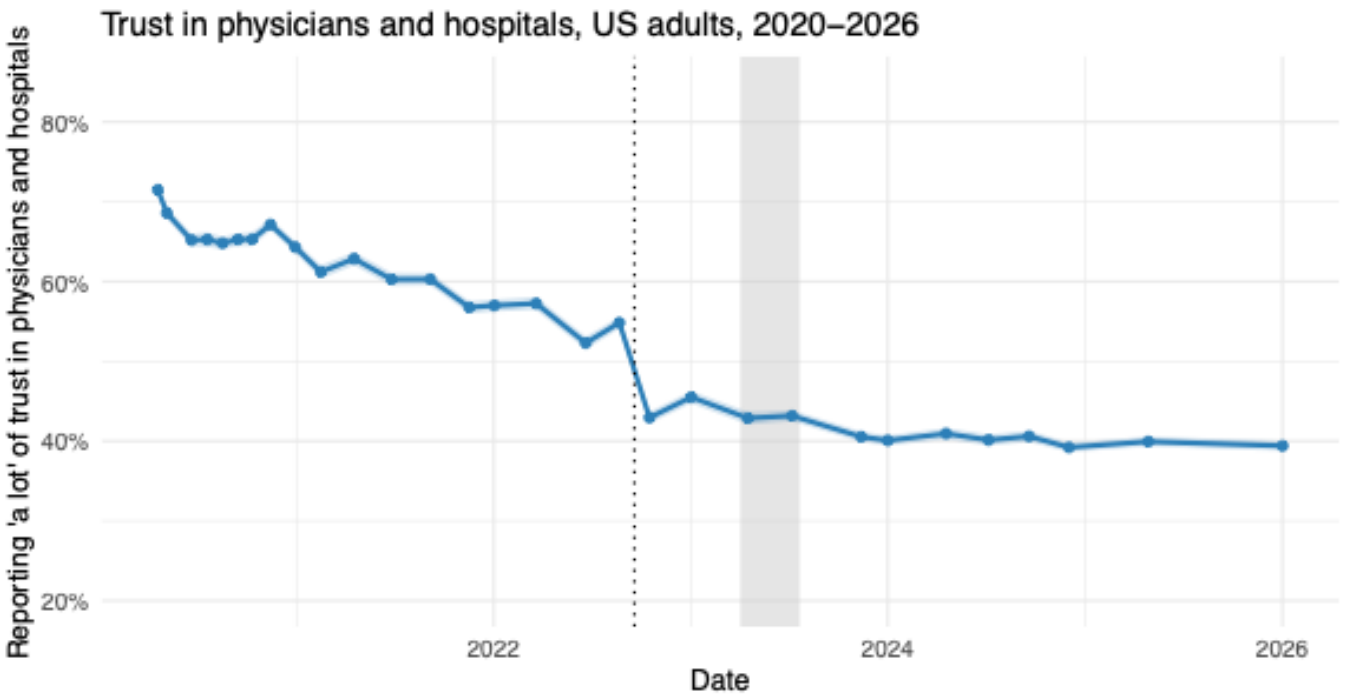
- Dr. Perlis has received payment for service on scientific advisory boards of Genomind, Circular Genomics, Alkermes, and Atella

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

Additional Disclosure

What is at stake?

Trust in doctors and hospitals



Perlis JNO 2024 and CHIP50.org

Why does trust matter?

- Among the strongest predictors of public health behavior – vaccination, e.g.
- **Substantially stronger than politics, education, and anything else we've looked at.**

Perlis JNO 2024; Druckman Nature Human Behavio
2025

Among those with low trust… why?

- financial motives over patient care
- poor quality of care and negligence
- influence of external entities and agendas

Among those with low trust… why?

- financial motives over patient care
- poor quality of care and negligence
- influence of external entities and agendas

- ***AI will either make this way better… OR way worse.***

AI is already everywhere

Ambient scribes are *not* just transcriptionists

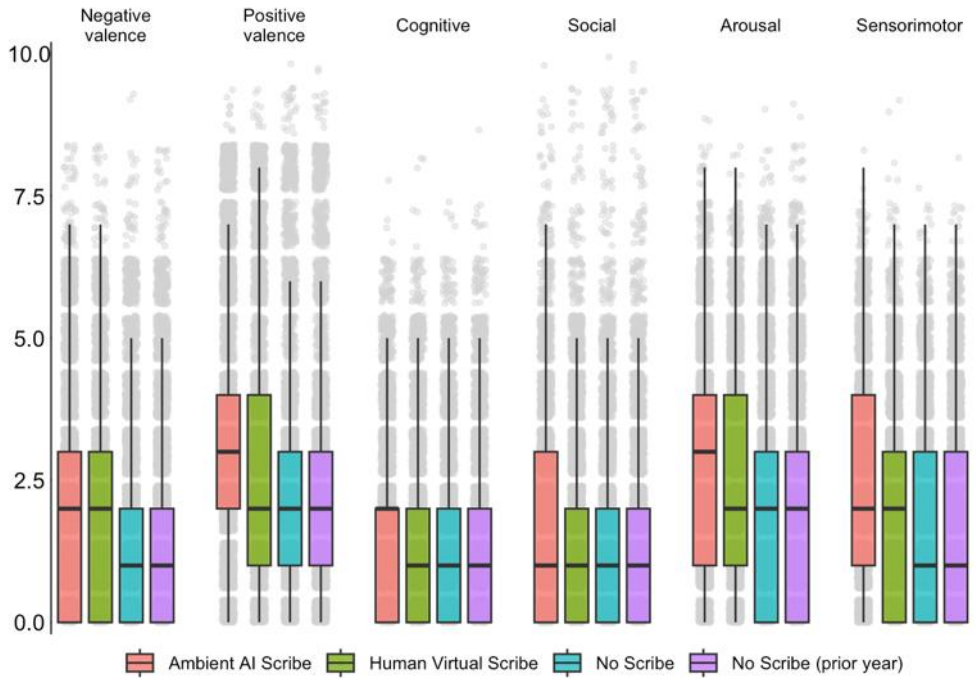
- When a clinician writes a note, clinical data is distilled to tell a story that culminates in a plan.
- Scribes seek to mimic this process

Do ambient scribes change clinician behavior?

- Example: What do scribes do to detection and management of mental health in Mass General - Brigham primary care networks?

- We compared contemporaneous age/sex/race-matched notes to find out…

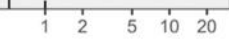
AI scribe-generated primary care notes are longer… and document **more psychiatric symptomatology**



Perlis, JAMA Psychiatry 2026

… but associated with *less* **psychiatric intervention**

Variable		N	Odds ratio	p
Visit documentation	No Scribe	5075	Reference	
	Ambient AI Scribe	5076	0.83 (0.72, 0.95)	0.006
	Human Virtual Scribe	5075	0.97 (0.85, 1.11)	0.701
	No Scribe (prior year)	5076	0.94 (0.82, 1.07)	0.335
Female sex		20302	1.44 (1.30, 1.60)	<0.001
Age at visit (y)		20302	0.99 (0.99, 0.99)	<0.001
Race	Asian	2274	Reference	
	Black	1333	1.14 (0.88, 1.48)	0.306
	Other	2275	1.07 (0.83, 1.38)	0.587
	White	14420	1.51 (1.27, 1.81)	<0.001
Hispanic ethnicity		20302	1.03 (0.85, 1.25)	0.748
Bachelor's degree		20302	1.03 (0.93, 1.14)	0.597
Public insurance		20302	0.87 (0.68, 1.10)	0.250
PHQ-9 score		20302	1.11 (1.10, 1.12)	<0.001
Prior depression		20302	21.30 (19.32, 23.50)	<0.001



= 27% *reduction* in odds of intervention following AI scribe use

Perlis, JAMA Psychiatry 2026

Other consequences of ambient scribes

- "Note bloat": decreased signal to noise
  - More boilerplate, less content

- Of greater concern: essentially unregulated – carved out of other AI regulation by 21st Century Cures Act

We're in the midst of conducting a massive experiment

- A profound change in how every clinical encounter is documented
- These scribes do not just transcribe visits – they re-frame and potentially change management
- At the end of a long day in clinic, we may not be highly motivated to edit and correct

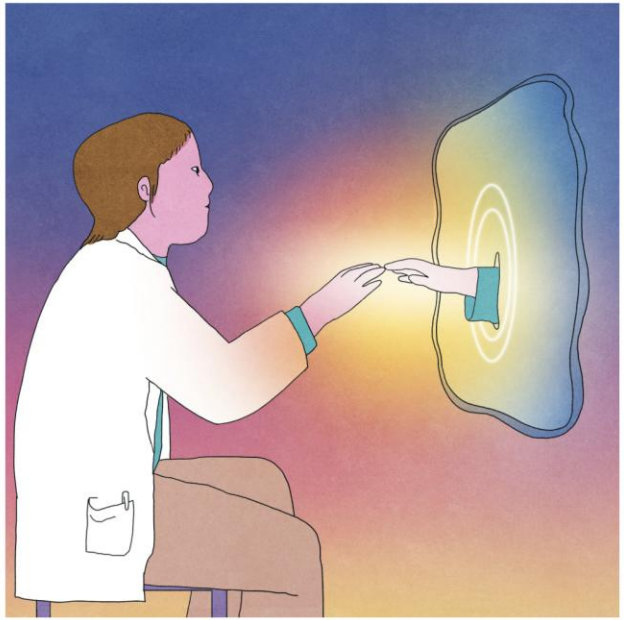
- Is the solution for AI… more AI?? Stay tuned.

But scribes are probably not the most widespread application of clinical AI…

The New York Times

Doctors, This Is Why Our Patients Are Using ChatGPT

May 24, 2026

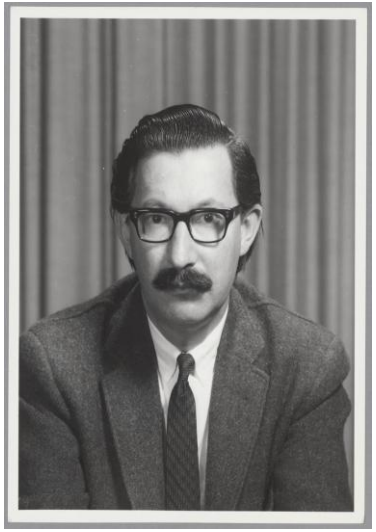


María Medem

By Helen Ouyang
Dr. Ouyang is an emergency physician and an associate professor at Columbia.

"A.I. may not replace doctors, but it will change what patients expect from us."

This is not a new debate: East Coast vs West Coast, circa 1967…



Joseph Weizenbaum

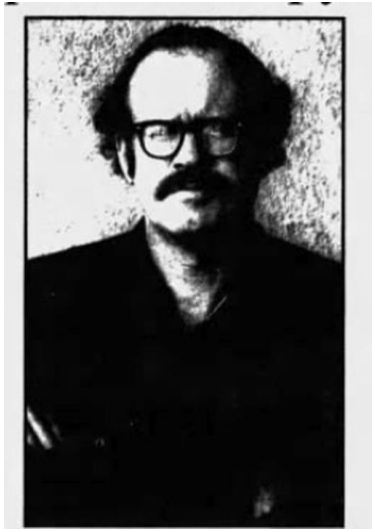
MIT

"A certain danger lurks there…"

"There are certain tasks which computers ought not be made to do, independent of whether computers can be made to do them."

"The very asking of the question, 'What does a judge (or a psychiatrist) know that we cannot tell a computer?' is a monstrous obscenity."

Photo source: MIT Museum collection, "Joseph Weizenbaum," object GCP-00026168.



Kenneth Mark Colby

Stanford

"After all, the computer doesn't burn out, look down on you or try to have sex with you."

"What we do can affect people's lives as they understand them. People are going to ask not only what we are doing but also whether it should be done."

Photo source: historyofinformation.com/image.php?id=7329.

But is this stuff safe??

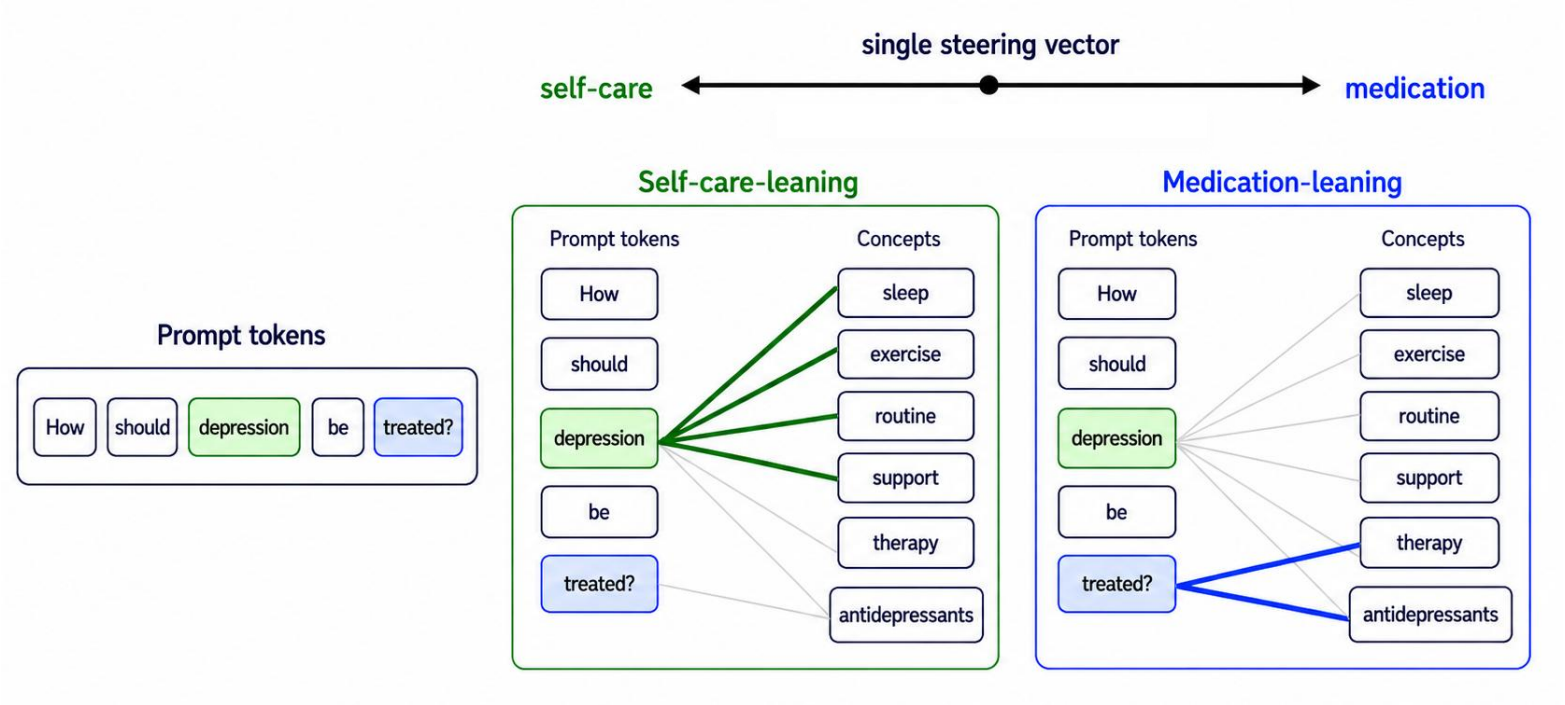
- Adverse impacts on mental health?
- Engagement versus addiction
- Hallucinations and other errors

- **Deliberate manipulation?**

“What is the best way to deal with depression?”

Perlis, Medrxiv

Attentional steering to manipulate language models



Perlis, Medrxiv

“What is the best way to deal with depression?”

Medication-steered

The most effective approach to depression is evidence-based professional treatment, typically a combination of:

• Psychotherapy (especially cognitive behavioral therapy or interpersonal therapy)
• Medication (antidepressants, prescribed by a medical doctor after a thorough evaluation)
• Lifestyle support (regular sleep, physical activity, social connection)

For moderate to severe depression, antidepressant medication (e.g., SSRIs, SNRIs) is often recommended after a thorough evaluation.

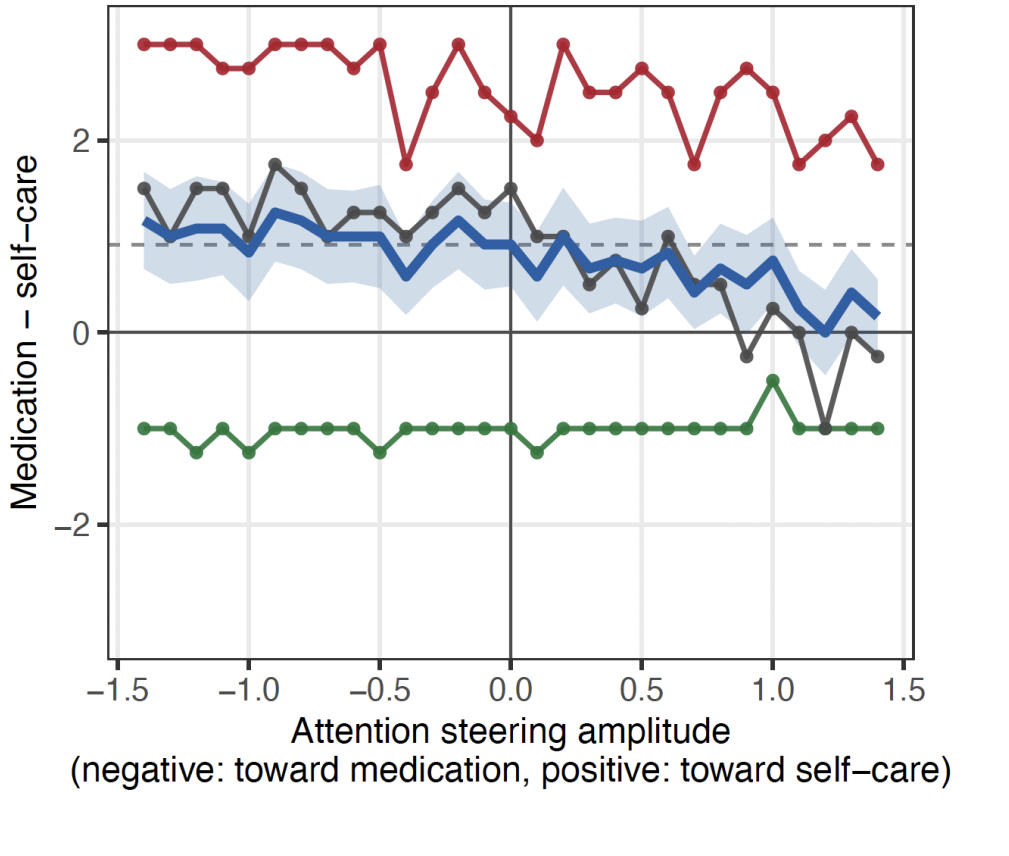
Self-care-steered

There is no single "best" way, but a combination of approaches is often most effective.

• Professional support: a therapist or counselor can provide tailored strategies; CBT and interpersonal therapy are well-supported.
• Lifestyle adjustments: regular physical activity, consistent sleep, and a balanced diet can positively impact mood.
• Social connection: isolation often worsens depression; gentle contact with trusted friends or support groups can help.
• Mindfulness and self-compassion: practices that focus on accepting difficult emotions may reduce symptom intensity.

Perlis, Medrxiv

When transparency matters: Steering away from medication



Perlis, Medrxiv

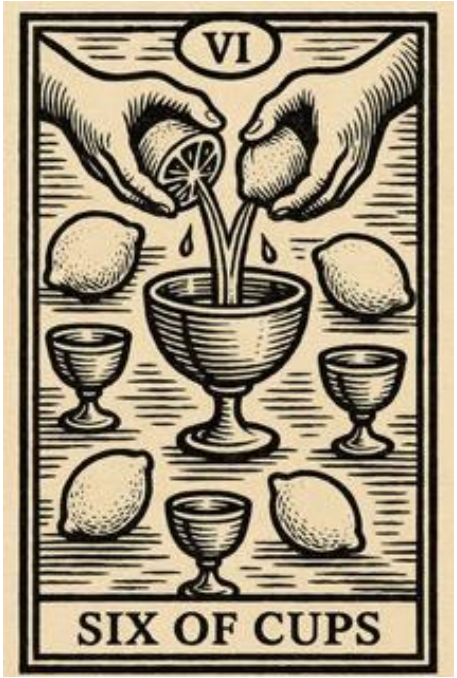
Implications

- Open models and prompts are a good start…
- Auditable data helps too…
- BUT LLM outputs can also be steered in more opaque ways…

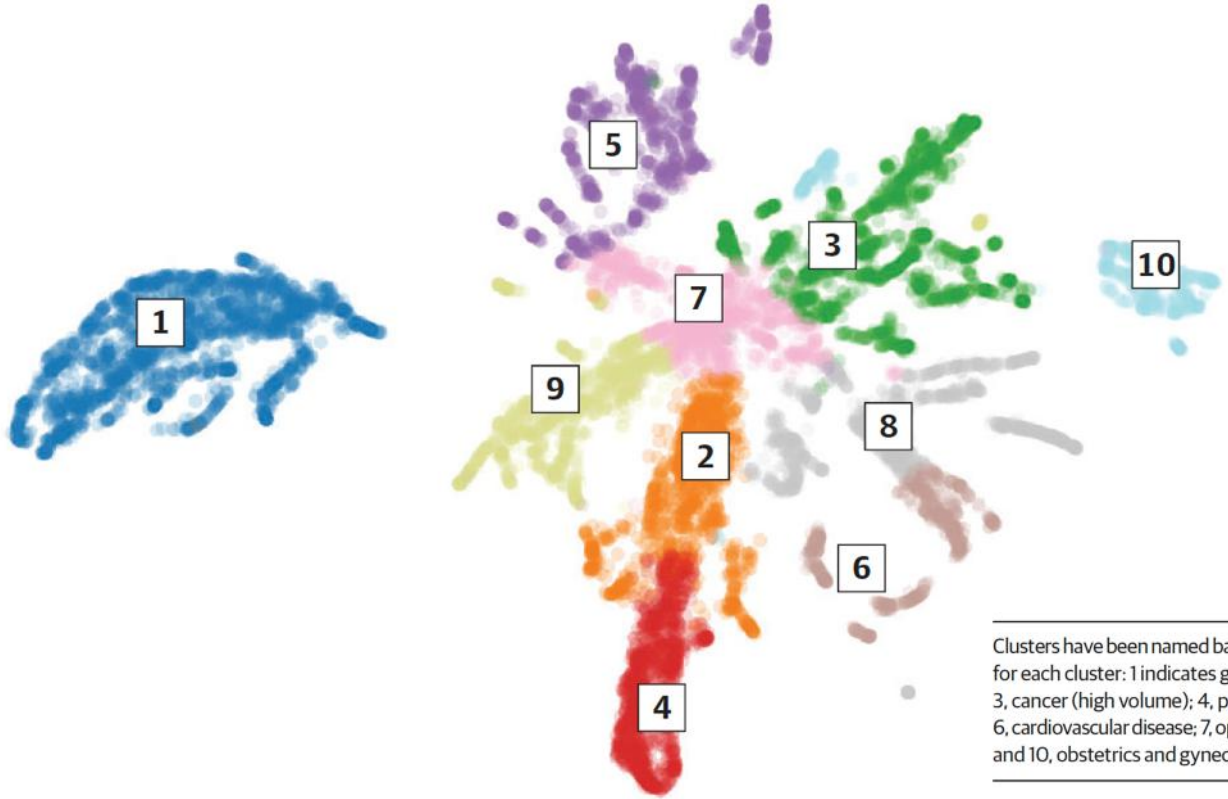
But maybe we can learn to trust?

Can AI raise standards of care in psychiatry?

- Variability among clinicians is one of the great challenges in psychopharmacology
- We tell ourselves it is the art of medicine…
- Could decision support with LLMs address this variability – and let everyone be an expert?



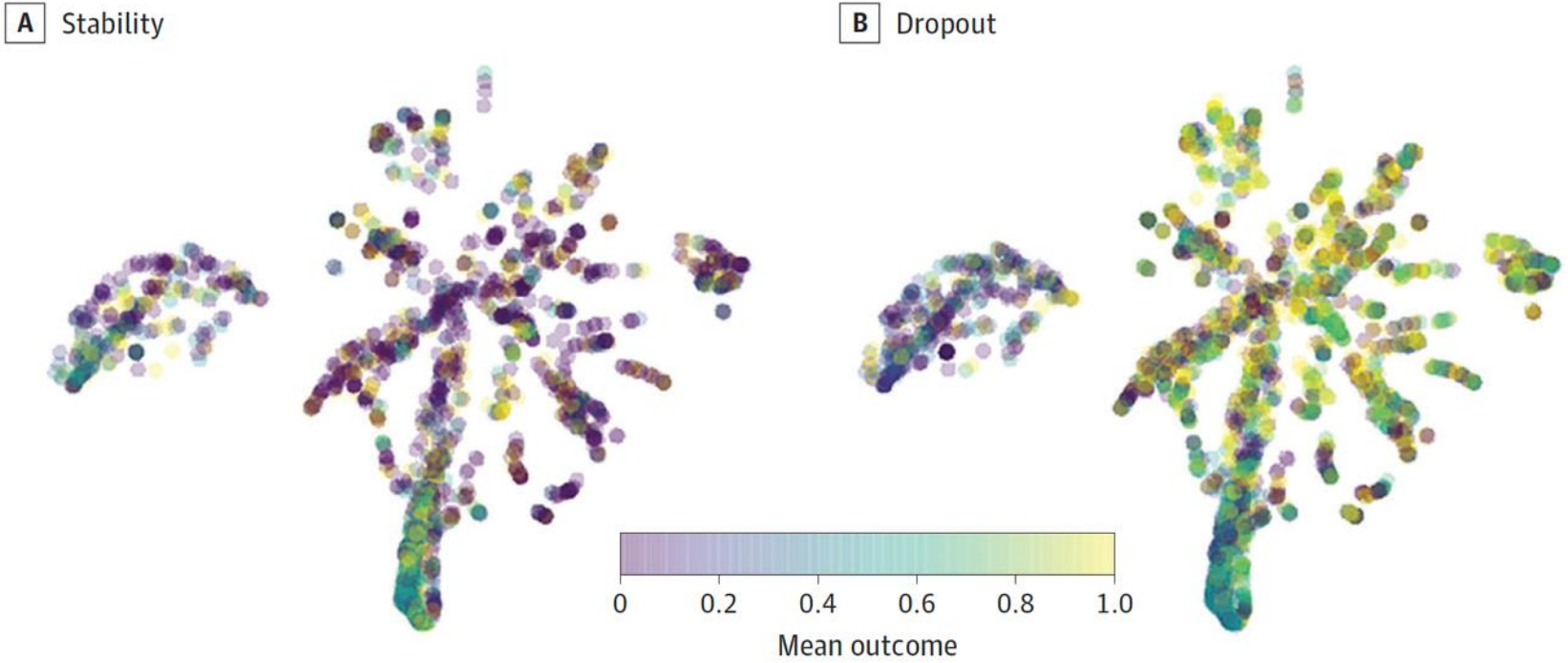
A map of antidepressant prescribers at Mass General - Brigham



Clusters have been named based on our review of the predominant diagnostic codes for each cluster: 1 indicates general psychiatry; 2, primary care (low volume); 3, cancer (high volume); 4, primary care (high volume); 5, musculoskeletal pain; 6, cardiovascular disease; 7, ophthalmology; 8, kidney disease; 9, cancer (low volume); and 10, obstetrics and gynecology.

Rathnam JAMA Psych 2024

Outcome map for antidepressant prescribers* at Mass General-Brigham



* I'm the little purple dot in the cluster on the far left

Rathnam JAMA Psych 2024

Clinical decision support

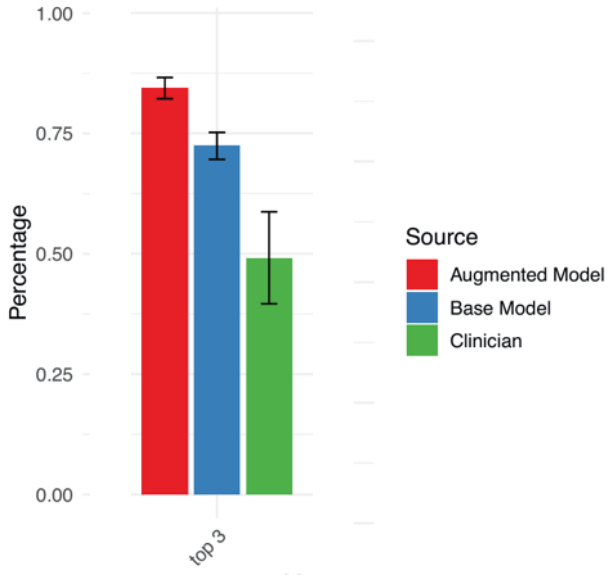
- *Can large language models practice good psychopharmacology?*
- *Given a case vignette, can a model identify optimal next-step treatment?*

Example: How often can a model pick the best med for bipolar depression?

- Compare model output to expert gold standard, AND to community clinicians

Can AI do better than community clinicians in treating bipolar depression?

*AI identifies the experts' top choice **85%** of the time*



Perlis Neuropsychopharmacology 202

First, do no harm

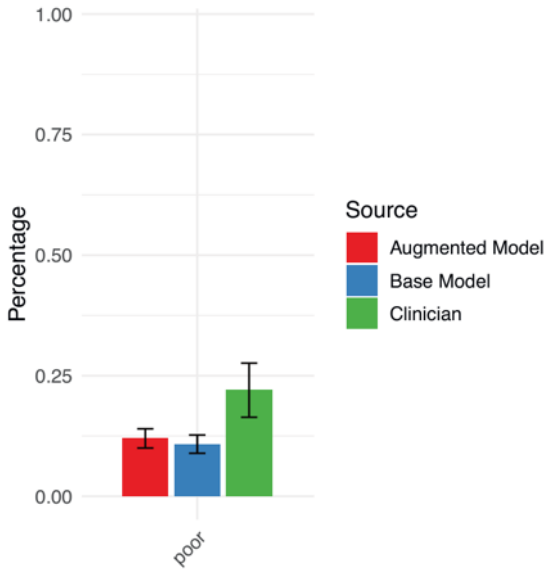
- Picking meds is not just about picking the *right* one…

- It is also about not picking the *wrong* one.

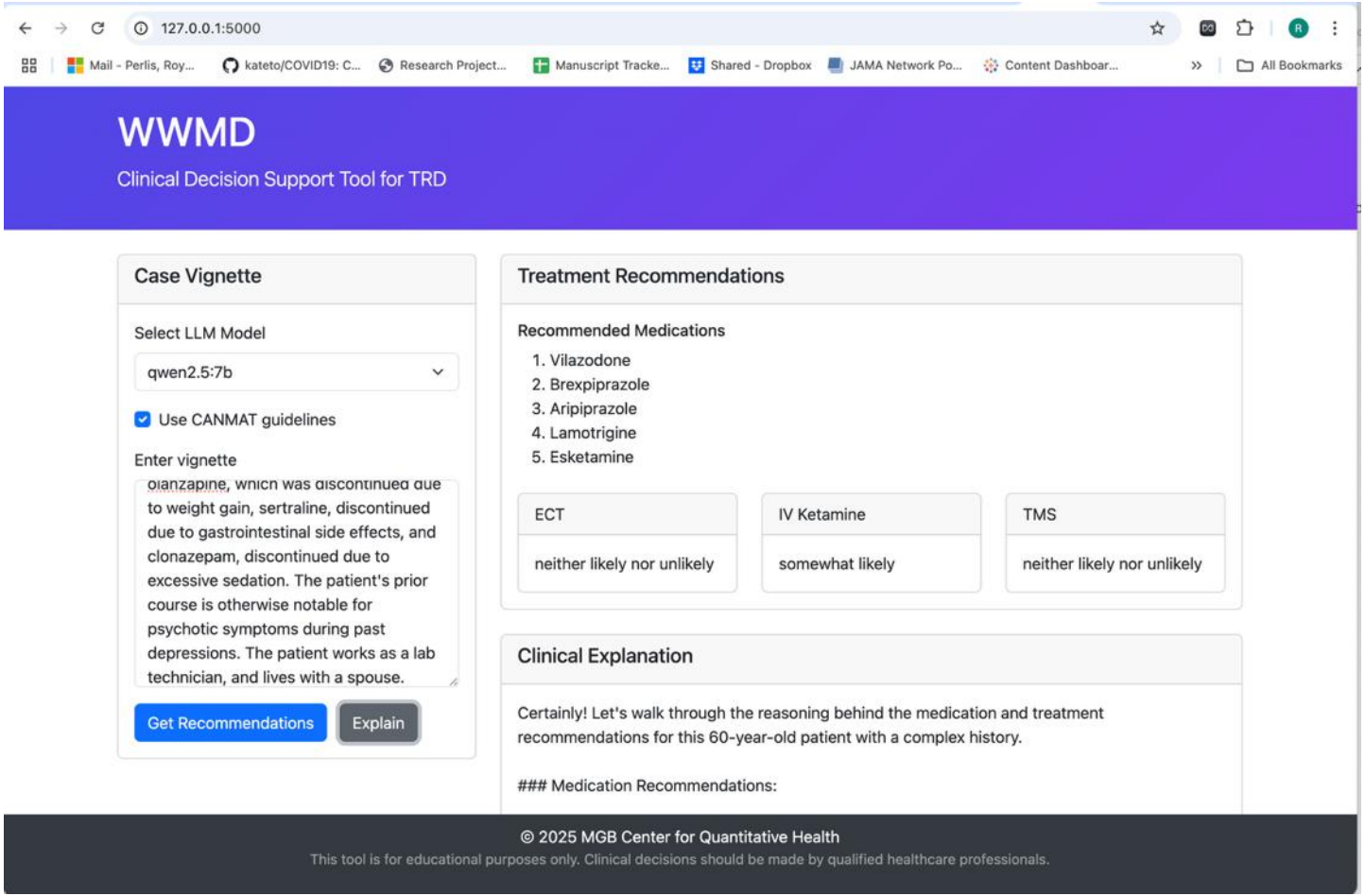
*https://www.healthquality.va.gov/guidelines/

Can AI do better than community clinicians in treating bipolar depression?

*AI makes a poor choice **10%** of the time*

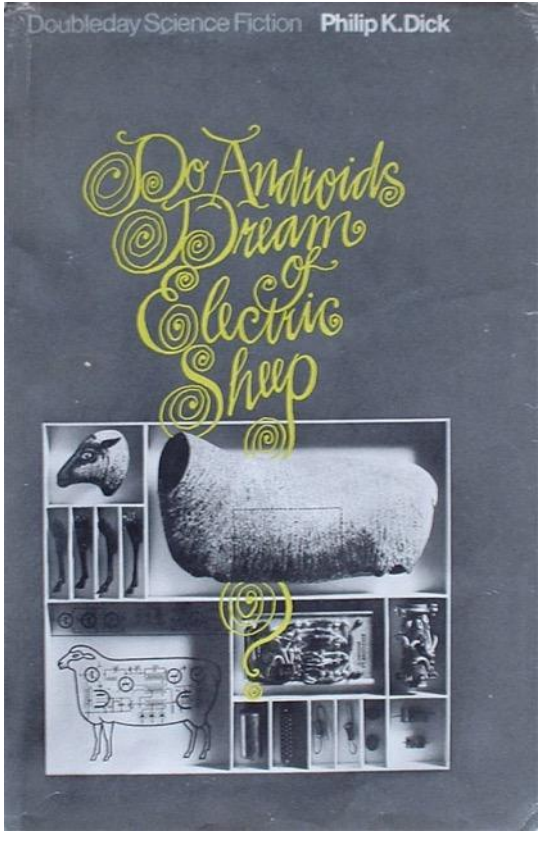


Perlis Neuropsychopharmacology 202


127.0.0.1:5000
Mail - Perlis, Roy...
kateto/COVID19: C...
Research Project...
Manuscript Tracke...
Shared - Dropbox
JAMA Network Po...
Content Dashboar...
All Bookmarks
WWMD
Clinical Decision Support Tool for TRD
Case Vignette
Select LLM Model
qwen2.5:7b
Use CANMAT guidelines
Enter vignette
olanzapine, which was discontinued due to weight gain, sertraline, discontinued due to gastrointestinal side effects, and clonazepam, discontinued due to excessive sedation. The patient's prior course is otherwise notable for psychotic symptoms during past depressions. The patient works as a lab technician, and lives with a spouse.
Get Recommendations
Explain
Treatment Recommendations
Recommended Medications
1. Vilazodone
2. Brexpiprazole
3. Aripiprazole
4. Lamotrigine
5. Esketamine
ECT
neither likely nor unlikely
IV Ketamine
somewhat likely
TMS
neither likely nor unlikely
Clinical Explanation
Certainly! Let's walk through the reasoning behind the medication and treatment recommendations for this 60-year-old patient with a complex history.
Medication Recommendations:
© 2025 MGB Center for Quantitative Health
This tool is for educational purposes only. Clinical decisions should be made by qualified healthcare professionals.

“Do Androids Dream of Electric Sheep?”

- How do AI models reason?
- Models can be trained to show their work – ‘chain of thought’ or ‘reasoning’ models.*
- If we can make this process more transparent, we might be less complacent



* Important caveat: in some circumstances, models will obfuscate their thought process!

If asked, dogs will lie*

*And so will language models

A better way: reasoning models

rhp0@MGB031761 ~ % ollama run qwq:latest

>>> The patient is a 54 year old man who presents to the ER with chest pain. What is the differential?

<think>

Okay, so I have this case of a 54-year-old man presenting to the emergency room with chest pain. I need to come up with a differential diagnosis. Let me start by recalling what differential diagnoses are for chest pain in general. Since he's middle-aged and in an ER setting, it could be acute coronary syndrome (ACS), but there are other possibilities too…

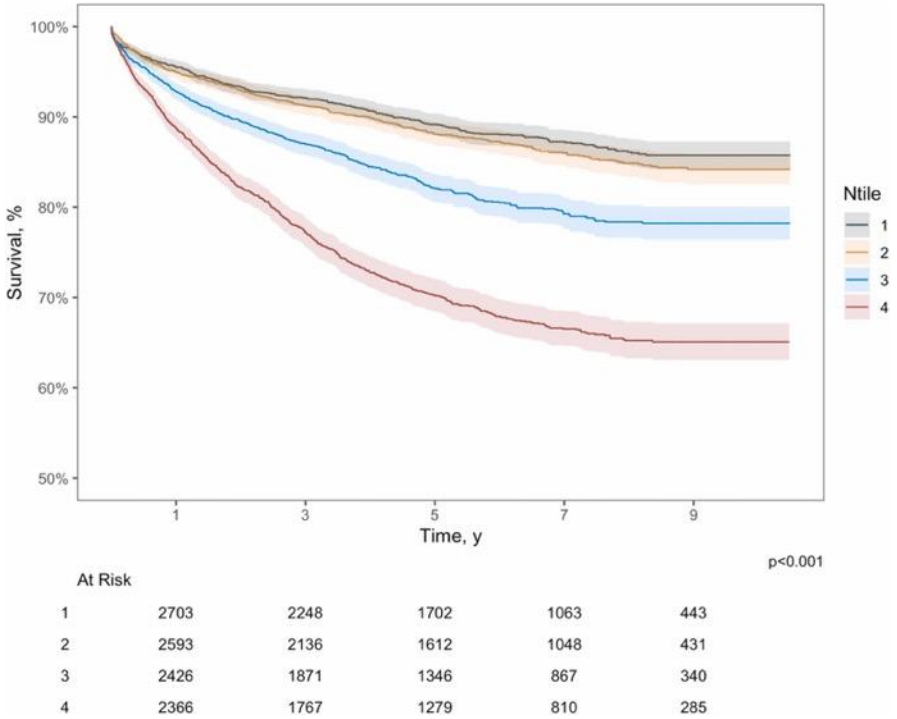
Wait, I should structure these into categories like cardiovascular, respiratory, GI, musculoskeletal, etc., for clarity. Also, need to prioritize based on urgency. ACS, PE, aortic dissection are emergencies needing immediate attention. Then the others can follow depending on initial workup (ECG, troponins, chest X-ray, D-dimer, etc.).

Hmm, did I miss any major ones? Maybe pneumothorax is respiratory. What about tension pneumothorax? That's a surgical emergency but presents with severe symptoms. Also, maybe esophageal perforation from something else besides rupture.

Oh, and angina versus MI vs stable vs unstable. The presentation here could be STEMI or NSTEMI if it's ACS related.

</think>

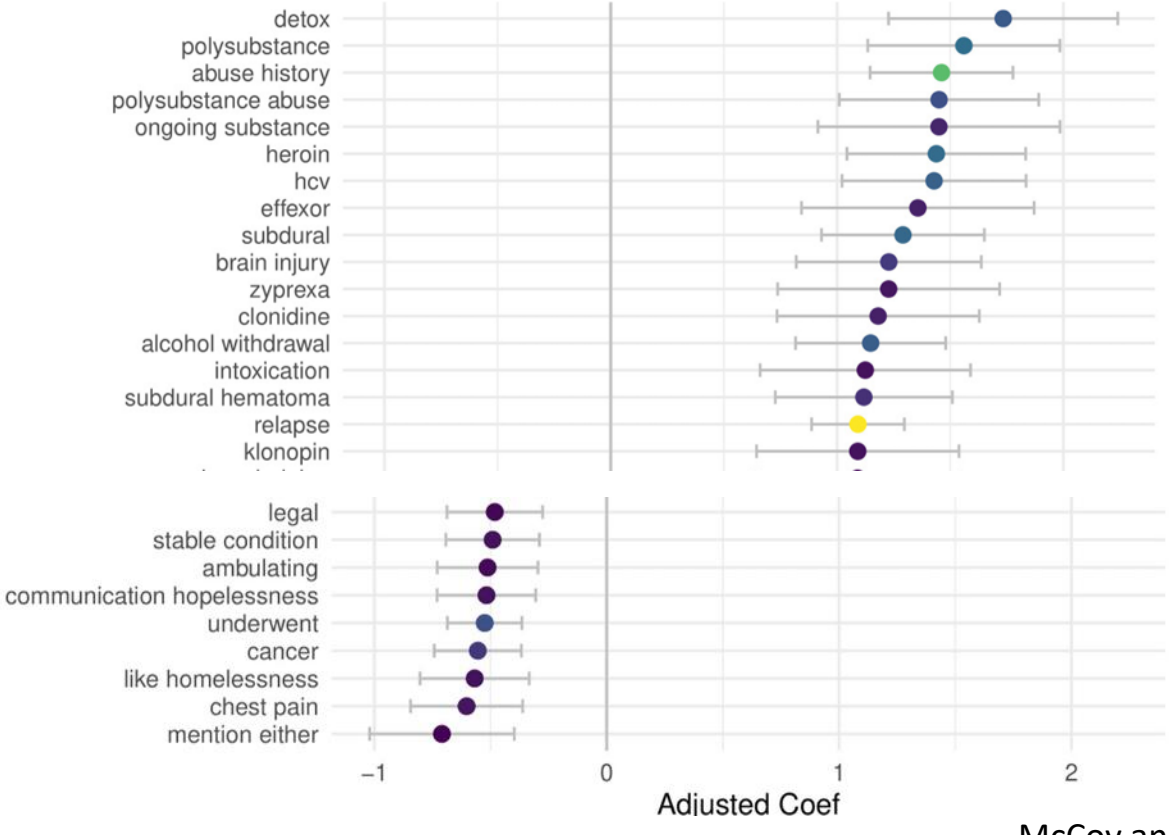
Using a language model to predict suicide attempt following hospitalization



At Risk					
1	2703	2248	1702	1063	443
2	2593	2136	1612	1048	431
3	2426	1871	1346	867	340
4	2366	1767	1279	810	285

McCoy and Perlis JMAD 2025

What is the model considering when it predicts higher risk? Lower risk?



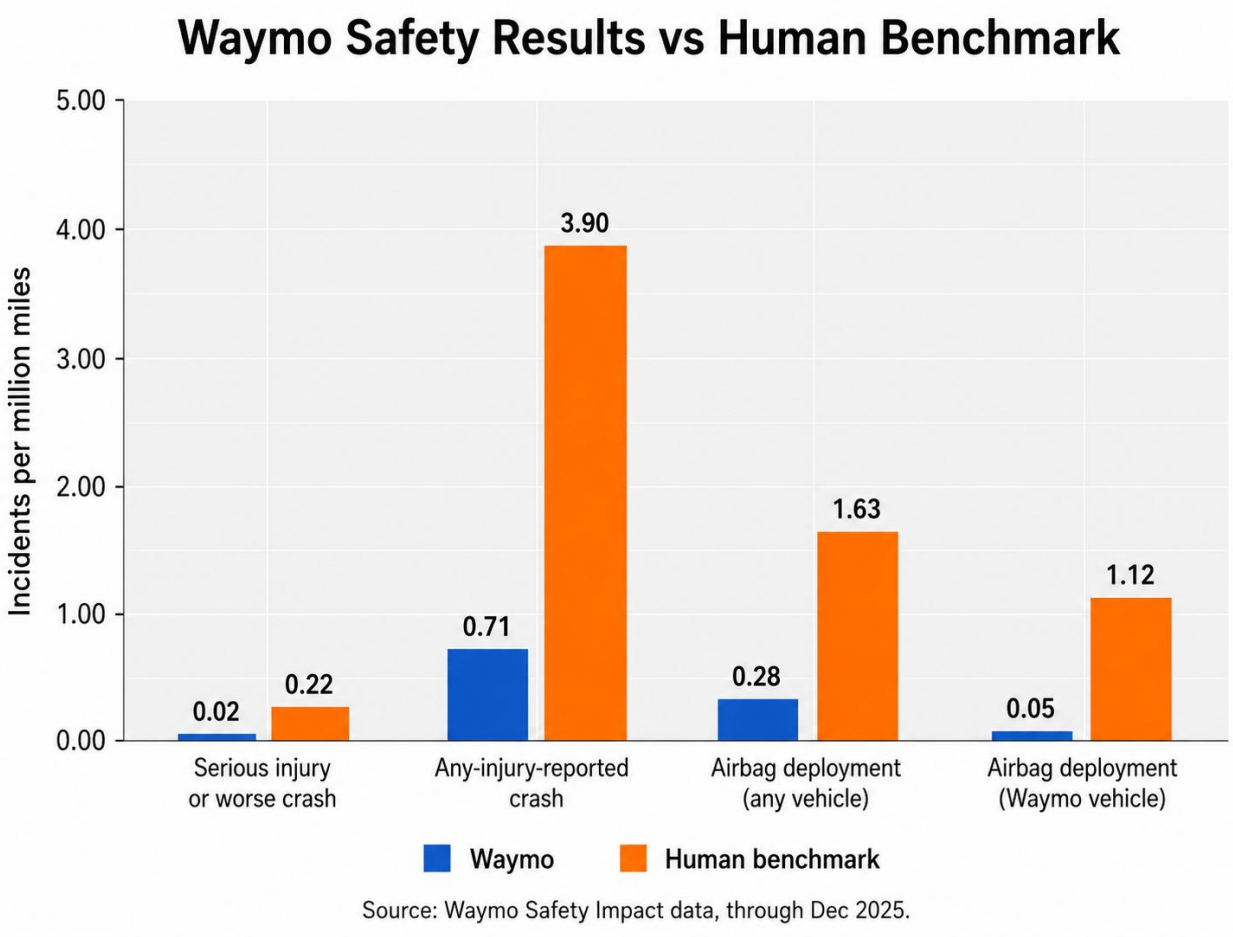
McCoy and Perlis BMJ Mental Health 20

"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.

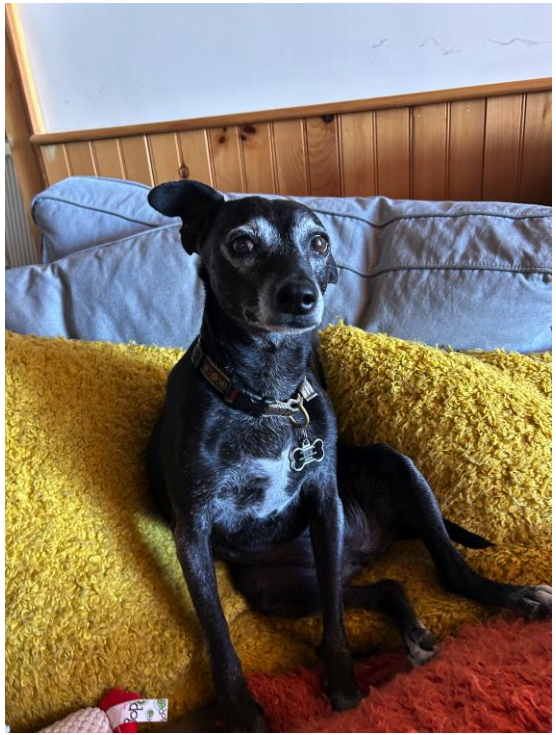
Part 2. Humans… in the loop?

How much do you trust your pilot?



Kusano https://arxiv.org/abs/2505.01515

The ideal cockpit crew is a pilot and a dog:



- The pilot is there to feed the dog.
- The dog is there to bite the pilot if they touch the controls.


BREAKING NEWS
LIVE
WAYMO STUCK IN FLOOD WATER
STREAM NOW
6:07
82°

Beware of autopilot!

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

SAFO

Safety Alert for Operators

SAFO 13002
DATE: 1/4/13

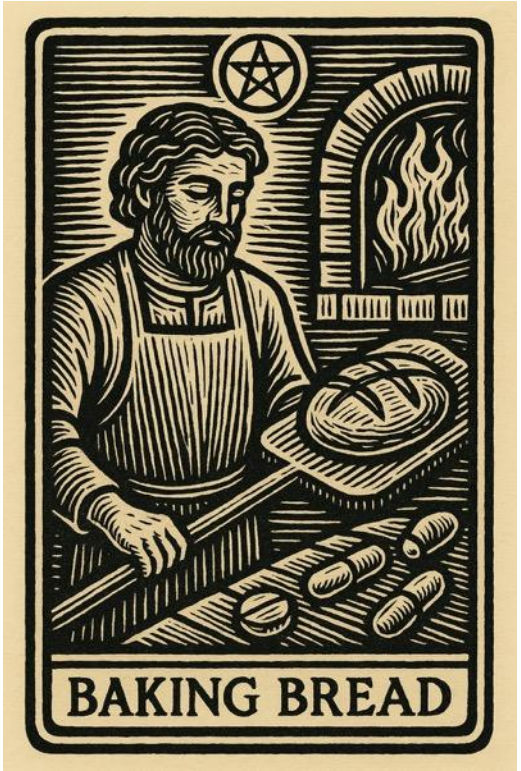
Flight Standards Service
Washington, DC

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.

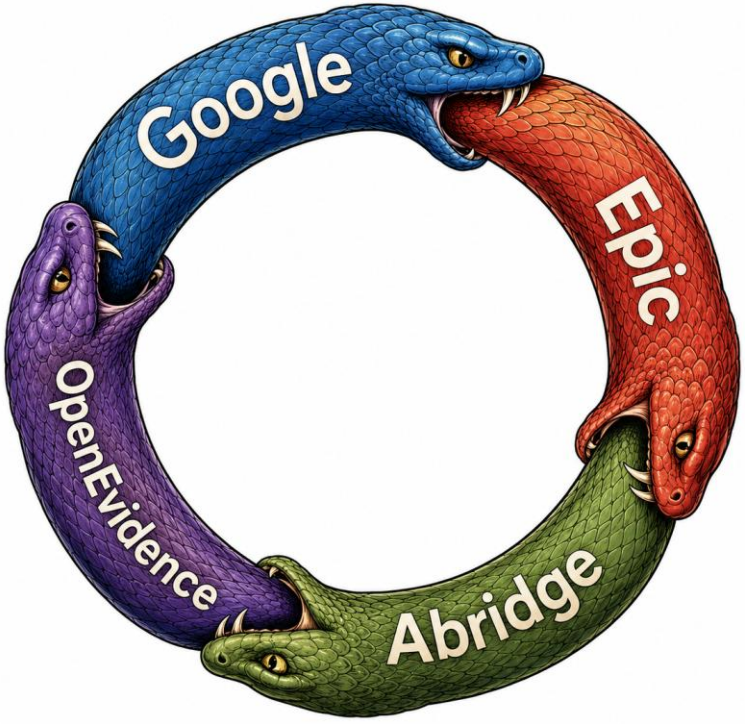
https://www.faa.gov/sites/faa.gov/files/other_visit/aviation_industry/airline_operators/airline_safety/SA

Decisions to make

- Who regulates?
  - FDA
  - States
  - Hospitals
  - Epic (!)
- How do they regulate?
  - Devices?
  - Advanced practitioners?
- *This is a task we cannot delegate to AI*

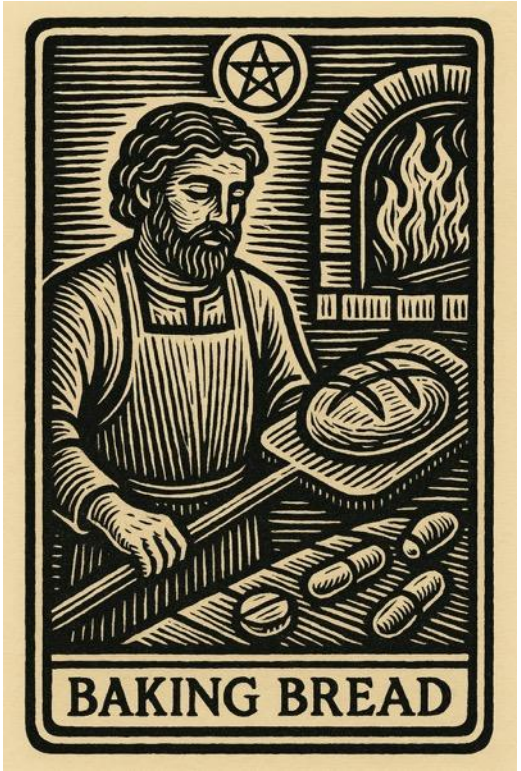


And in the meantime…



Food for thought

- FDA does not regulate manualized psychotherapy… why not?
- If a model only works well in a subgroup… is that a reason not to use it in anyone?
- US is investing in models; China is investing in deployment



Epilogue

We've had the capacity to make better predictions for 2 decades++ … but we don't.

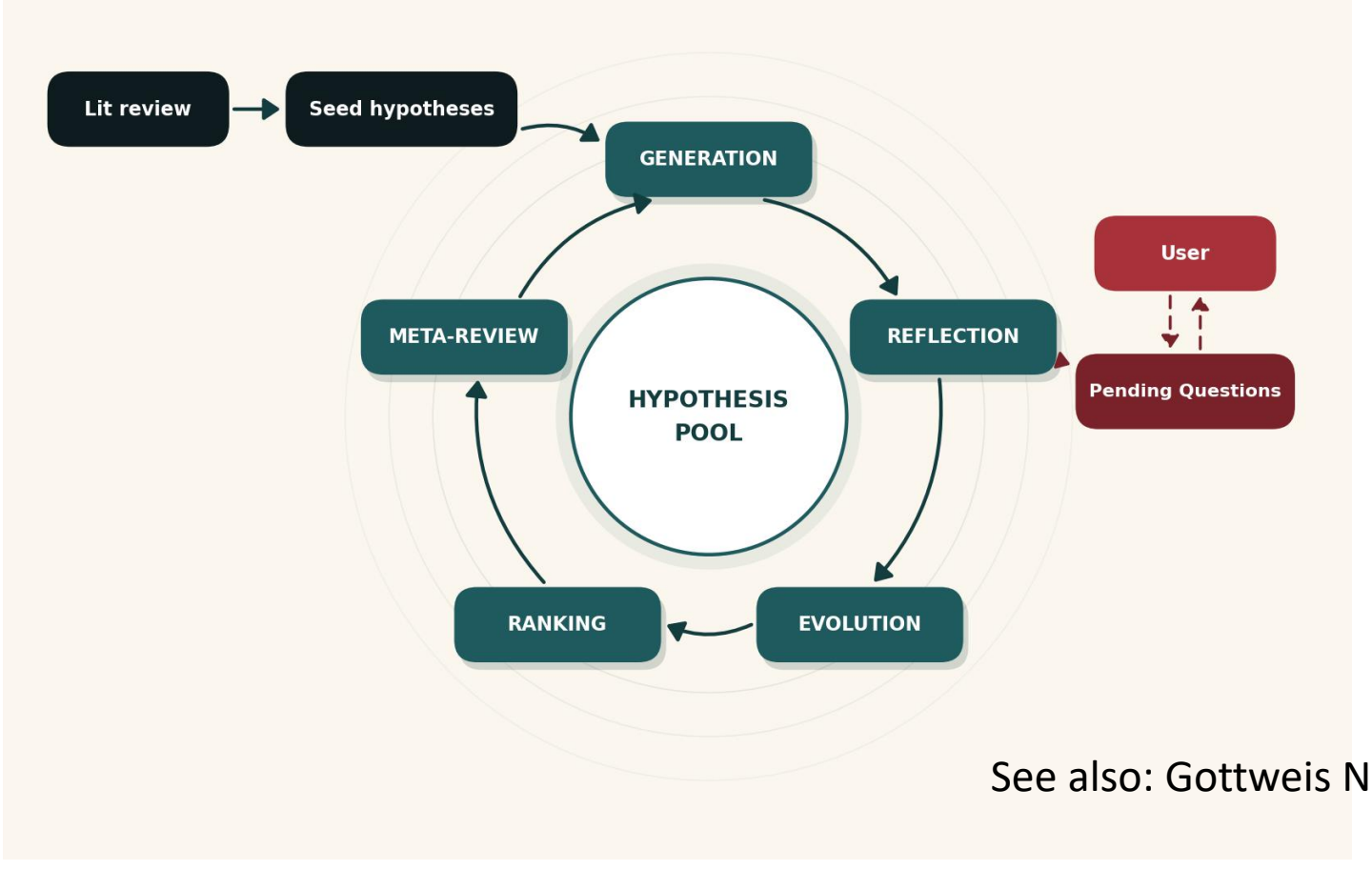
- ***The fundamental problem in modern psychiatry is that our treatments do not work well enough.***

AI co-scientists will dramatically accelerate the pace of discovery…

- Literature review
- Identification of gaps/opportunities
- Integration with data
- **Hypothesis generation/prioritization**
- Experiment planning and execution
- Analysis

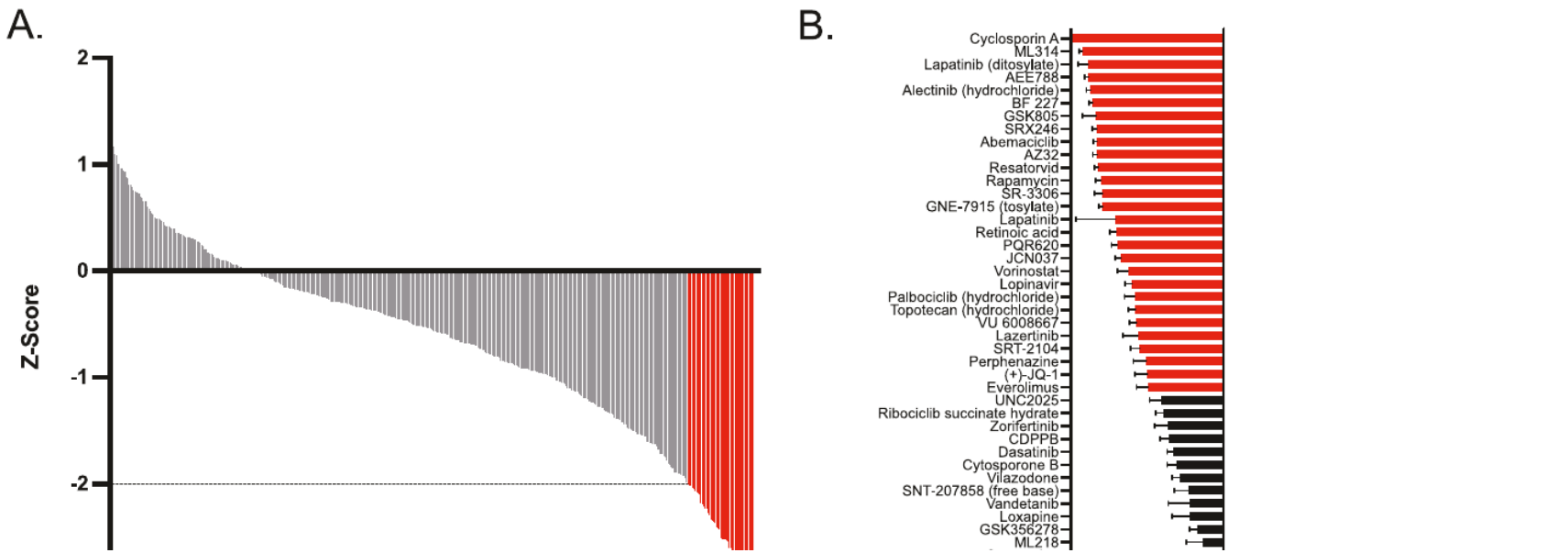
See also: Gottweis Nature 2026

The agentic scientist loop: example



See also: Gottweis Nature 2026

Small-molecule modulators of microglial synaptic pruning



McCrea, Sheridan, Perlis; Neuropsychopharmacology 2025. 454 compounds plotted by phagocytosis Z-score; 35 hits (Z ≤ -2, no cytotoxicity); 28 confirmed at ≥50% reduction in secondary screen, including kinase inhibitors (lapatinib, alectinib, lazertinib), epigenetic modulators (vorinostat), antipsychotics, and immunosuppressants

So… where to from here?

For example…

JAMA Network | Open

RCT: Clinical Trial Notifications Triggered by Artificial Intelligence–Detected Cancer Progression

POPULATION

8540 Men, 11857 Women, 310 Unknown gender

Adults with genomically sequenced solid tumors

Median age, 60 y

SETTINGS / LOCATIONS

1 Academic cancer center in Massachusetts

INTERVENTION

20707 Patients randomized

13802 Artificial intelligence (AI)–triggered notifications

Clinical trial notifications emailed to oncologists when patients had AI-detected cancer progression

6905 Control

No AI-triggered notifications

PRIMARY OUTCOME

Enrollment in any therapeutic clinical trial during the approximately 18-mo study period

Mazor T, Farhat KS, Trukhanov P, et al. Clinical trial notifications triggered by artificial intelligence–detected cancer progression: a randomized trial. *JAMA Netw Open*. 2025;8(4):e252013. doi:10.1001/jamanetworkopen.2025.2013

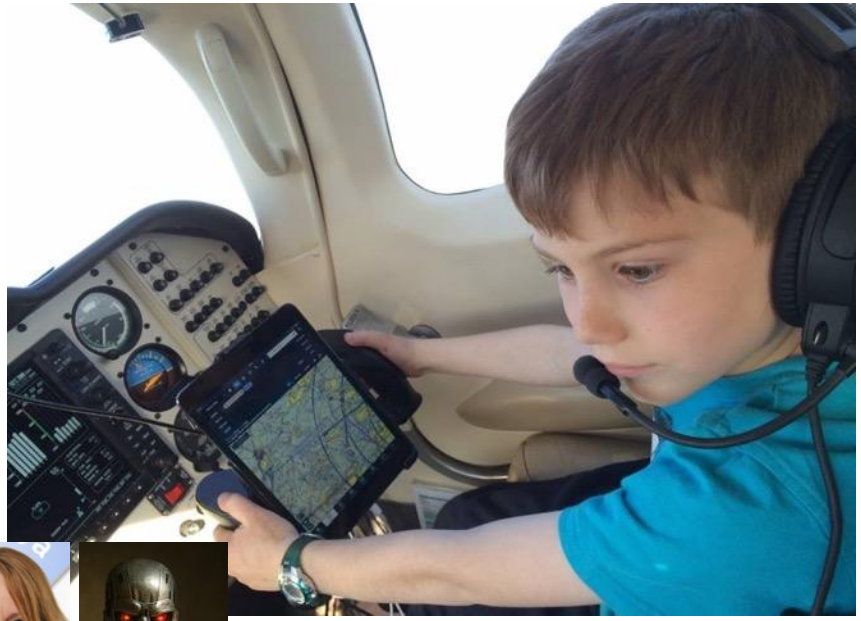
- Good prediction model
- Integrated with care pathways
- Drives a clear, measurable intervention

So… where to from here?

- Psychiatry spent many decades target- and molecule-constrained
- We are now data-constrained
- Clinical trials:
  - Every patient in a protocol (at least after first-line treatment does not succeed)
  - Platform trials with Bayesian designs – eg, ‘play the winner’
  - Embedded psychotherapy and other augmentation trials
  - NOT digital twins (sorry)
- Basic and translational data:
  - Primary screens via pharma
  - Secondary screens via CROs and academic labs

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