Consequential decision making and general-purpose AI

Suresh Venkatasubramanian
Brown University

high stakes decision making affects people’s

- lives
- opportunities
- access to services

high stakes decision making affects people's

- lives
- opportunities
- access to services

And so we want the process to be

- safe and effective
- transparent and accountable
- equitable

automated decision systems (ADSs) before 2023

automated decision systems (ADSs) before 2023

…are **task specific** (“predict likelihood that a child will be placed in foster care two years after the first report”)

automated decision systems (ADSs) before 2023

…are **task specific** ("predict likelihood that a child will be placed in foster care two years after the first report")

…are usually built with custom data for the task

automated decision systems (ADSs) before 2023

…are **task specific** (“predict likelihood that a child will be placed in foster care two years after the first report”)

…are usually built with custom data for the task

…have knobs (parameters) that are usually **interpretable** in the context of the domain (“number of times a parent has had a mental health check-in in the last 365 days”)

automated decision systems (ADSs) before 2023
can be evaluated

automated decision systems (ADSs) before 2023
can be evaluated

…with respect to a specific target (“how many children did actually end up in foster care after two years”)

automated decision systems (ADSs) before 2023
can be evaluated

…with respect to a specific target ("how many children did actually end up in foster care after two years")

…with an understanding of type and nature of inputs

automated decision systems (ADSs) before 2023
can be evaluated

…with respect to a specific target (“how many children did actually end up in foster care after two years”)

…with an understanding of type and nature of inputs

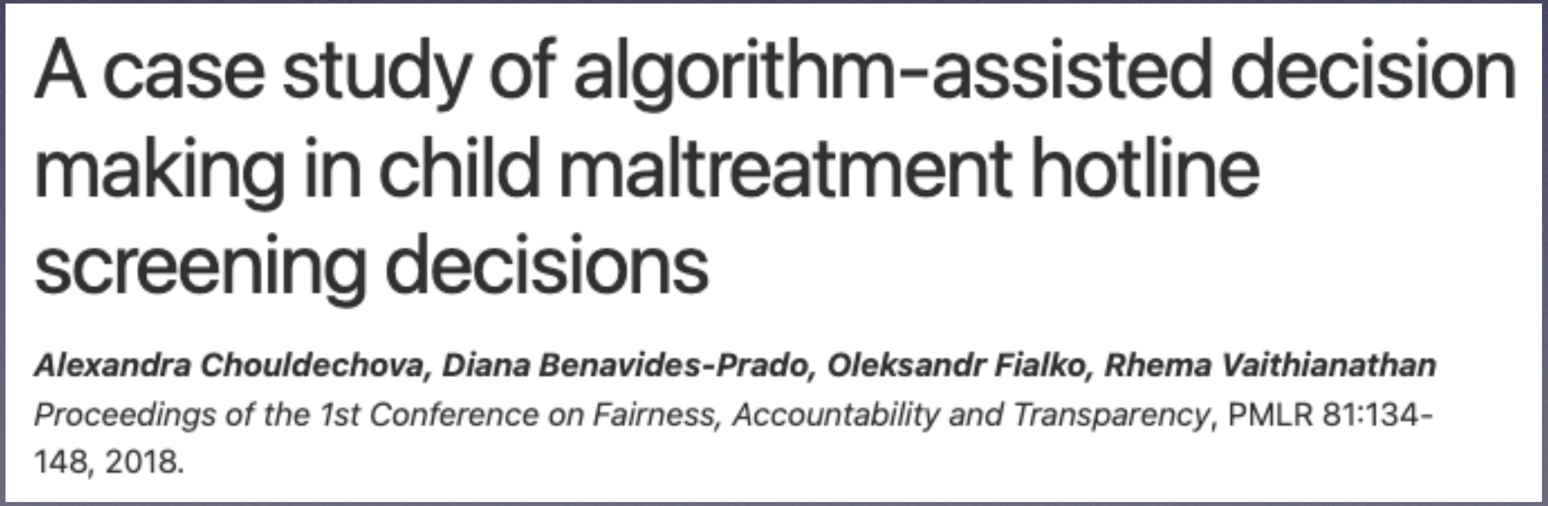
…in a reproducible way

automated decision systems (ADSs) before 2023
can be evaluated

…with respect to a specific target ("how many children did actually end up in foster care after two years")

…with an understanding of type and nature of inputs

…in a reproducible way



A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions

Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, Rhema Vaithianathan

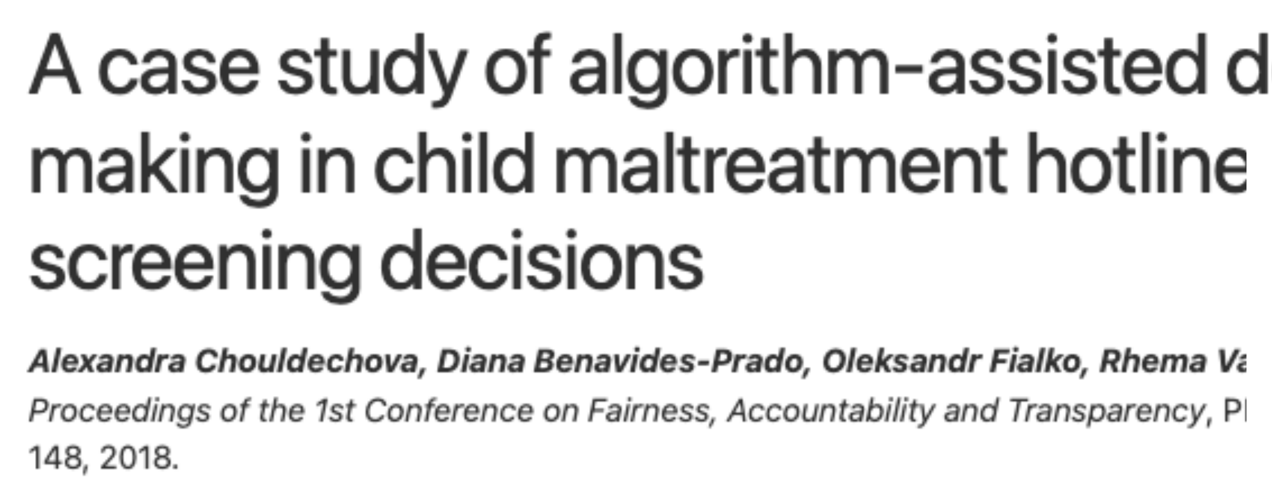
Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:134-148, 2018.

automated decision systems (ADSs) before 2023
can be evaluated

…with respect to a specific target (“how many children did actually end up in foster care after two years”)

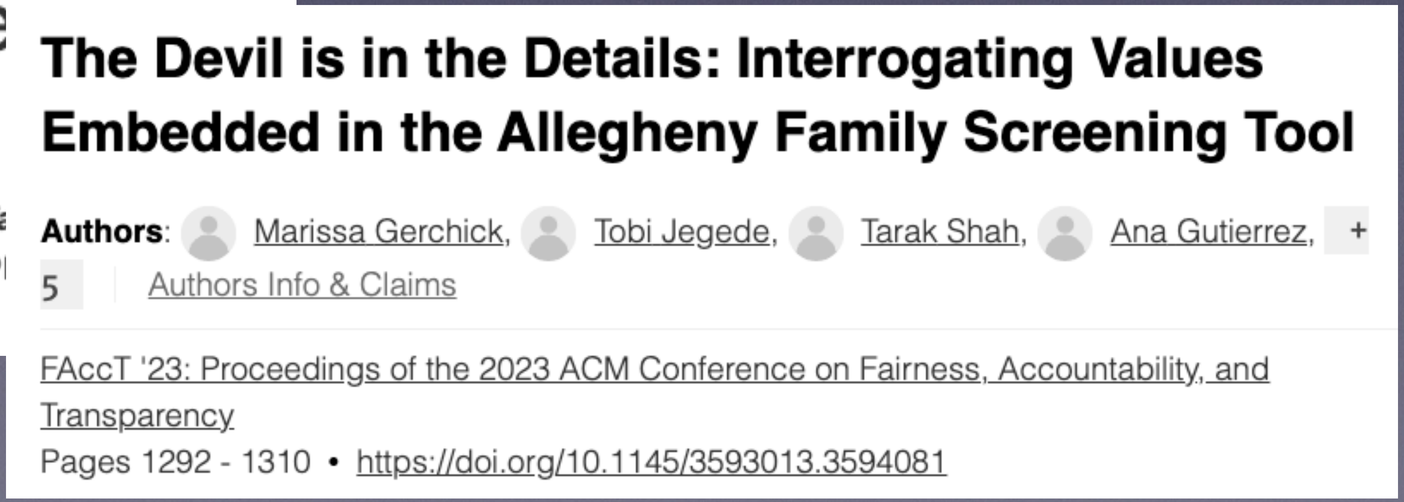
…with an understanding of type and nature of inputs

…in a reproducible way



A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions

Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, Rhema Va
Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Pl
148, 2018.



The Devil is in the Details: Interrogating Values Embedded in the Allegheny Family Screening Tool

Authors: Marissa Gerchick, Tobi Jegede, Tarak Shah, Ana Gutierrez, + 5 | Authors Info & Claims

FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency
Pages 1292 - 1310 • https://doi.org/10.1145/3593013.3594081

automated decision systems (ADSs) before 2023
have **many** problems

automated decision systems (ADSs) before 2023
have **many** problems

but there are ways to decide go/no-go or how to mitigate …

BLUEPRINT FOR AN AI BILL OF RIGHTS

MAKING AUTOMATED SYSTEMS WORK FOR THE AMERICAN PEOPLE

OSTP

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

Aequitas
Bias & Fairness Audit

VerifyML

Validating input features, testing for data quality, bias mitigation, building governance frameworks…

What happens when we move to general-purpose AI tools?

general purpose AI tools (post 2023)

...are trained to **converse** in natural language

... are trained to be **generally useful**

...... generate human-like text and styles

...... process and generate images

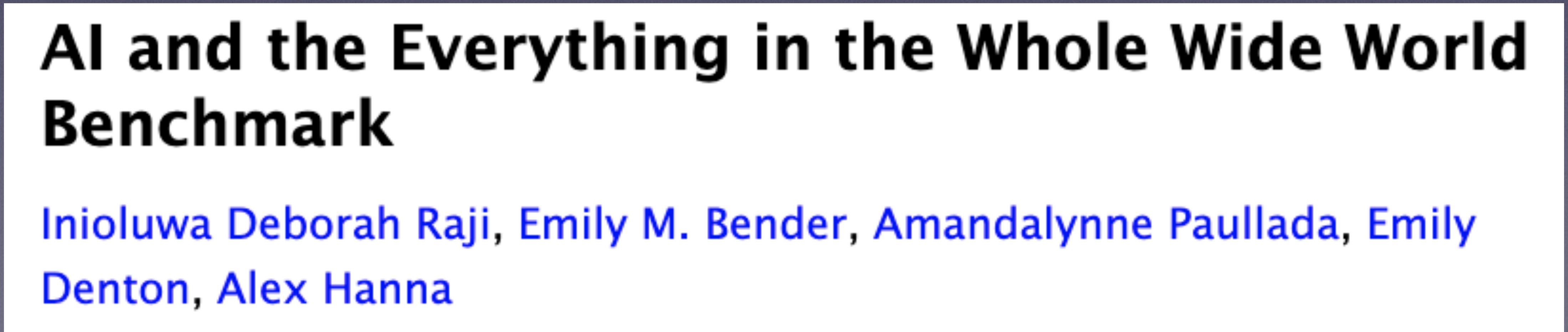
...... perform “reasoning” tasks

...... write and debug code

... often within the same system

general purpose AI tools (post 2023) **cannot** be evaluated

…with respect to a specific task

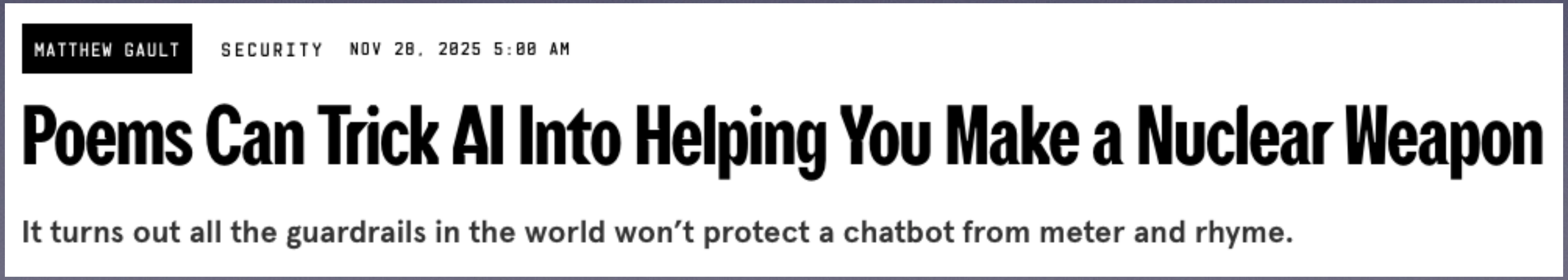


AI and the Everything in the Whole Wide World Benchmark

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, Alex Hanna

general purpose AI tools (post 2023) **cannot** be evaluated

…with an understanding of type and nature of inputs



MATTHEW GAULT SECURITY NOV 20, 2025 5:00 AM

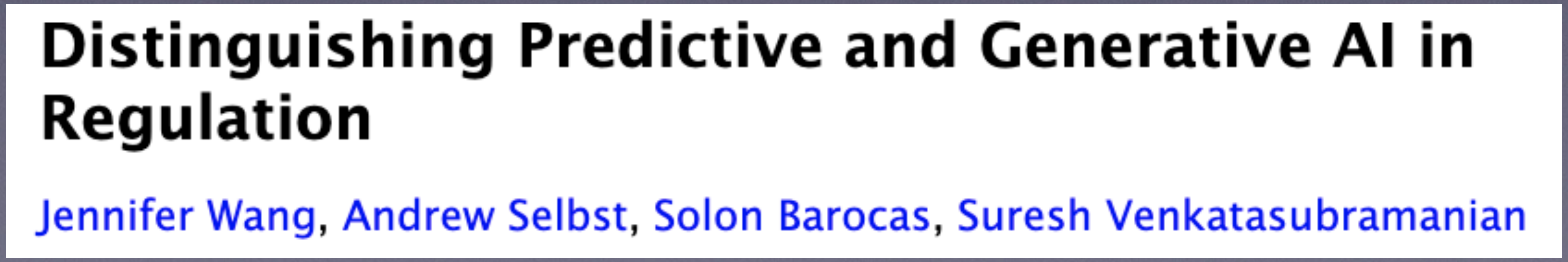
Poems Can Trick AI Into Helping You Make a Nuclear Weapon

It turns out all the guardrails in the world won't protect a chatbot from meter and rhyme.

general purpose AI tools (post 2023) **cannot** be evaluated

…reproducibly

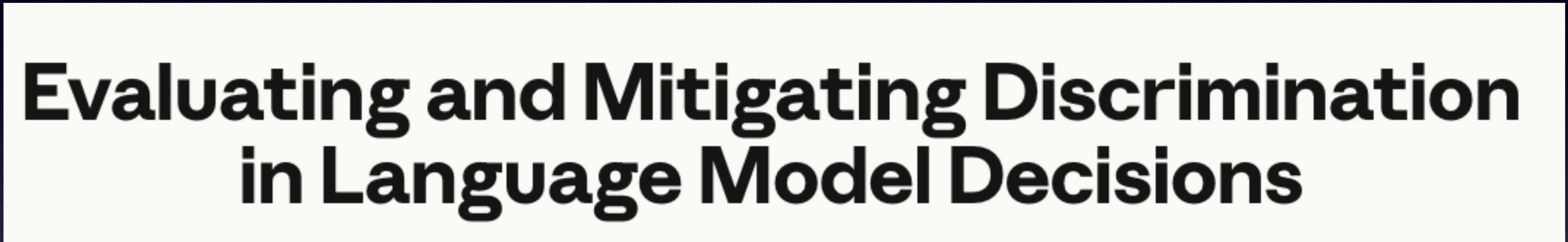
“[GPAI] is built on a highly distributed value chain that complicates accountability.”



Distinguishing Predictive and Generative AI in Regulation

Jennifer Wang, Andrew Selbst, Solon Barocas, Suresh Venkatasubramanian

general purpose AI tools (post 2023) **cannot** be evaluated



Evaluating and Mitigating Discrimination in Language Model Decisions

To get the model to not discriminate, it was asked
… “don’t discriminate”
… “really don’t discriminate”
… “really really don’t discriminate”
… “really really really really don’t discriminate”

general purpose AI tools (post 2023) are very complex

COMPAS used

137 parameters

GPT3 had

175,000,000,000 parameters

GPT4 probably has

1,750,000,000,000 parameters

general purpose AI tools (post 2023) **are very complex**

COMPAS parameter: “what is the zip code where the individual lives”

LLM parameter: “weight for attention head 23 in layer 54

LLM parameters are not usefully interpretable

Can we just ask the LLM to give us a rationale?

LLMs can “reason” using “chain of thought” prompting.

But….

… they can still hallucinate a response

… the response can be post hoc

… even the response needs validation…

… the response is often misleading

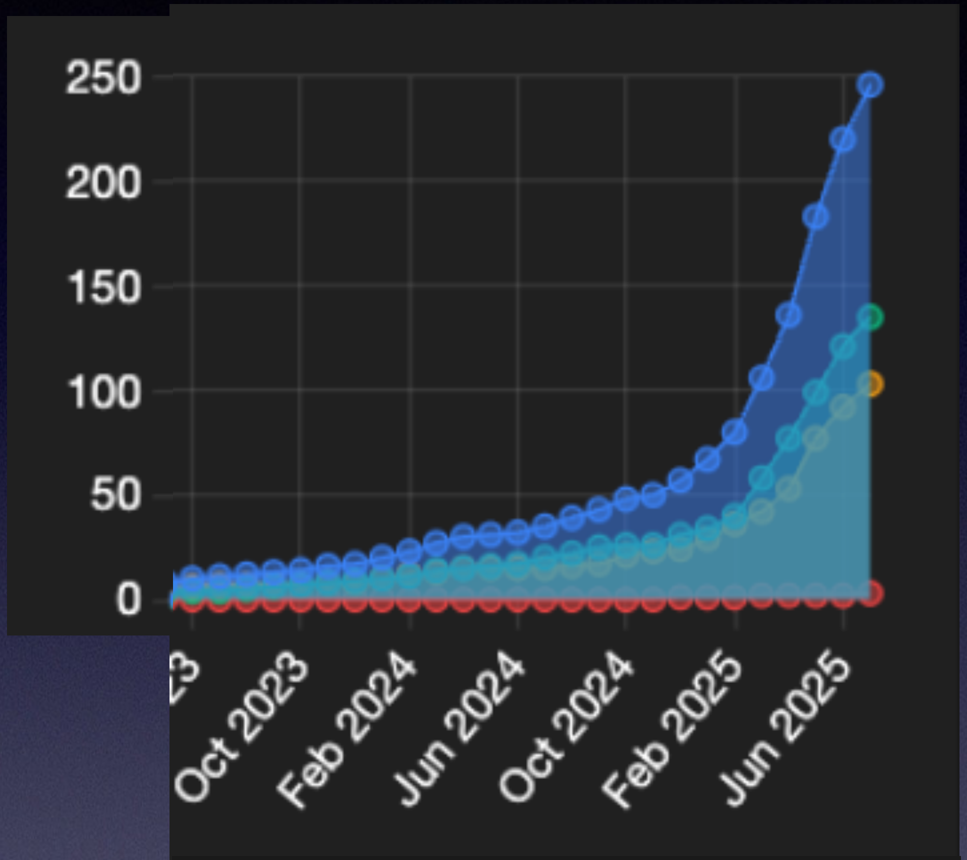
So when CAN we use general purpose AI tools?

… when the stakes are low (*document summaries)*

… the output can be independently verified (*coding)*

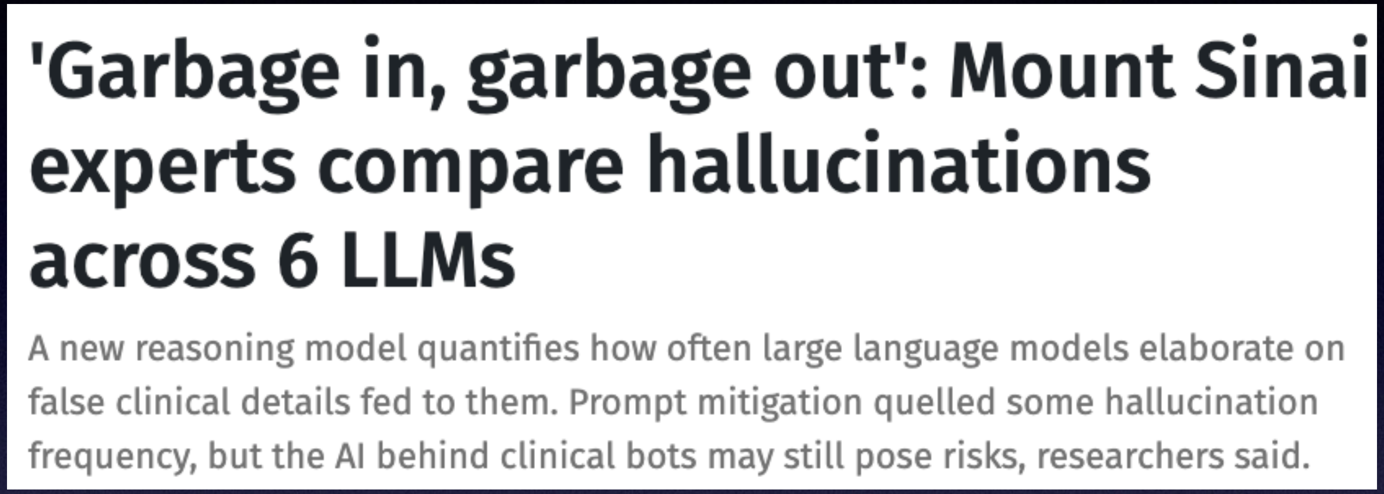
… there’s a broader accountability framework (*drug discovery)*

But even then we have to be careful…



Tracking Fake Citations in Court (e.g., AI Hallucinations) (246 cases)

https://www.polarislab.org/ai-law-tracker.html



'Garbage in, garbage out': Mount Sinai experts compare hallucinations across 6 LLMs

A new reasoning model quantifies how often large language models elaborate on false clinical details fed to them. Prompt mitigation quelled some hallucination frequency, but the AI behind clinical bots may still pose risks, researchers said.

So should we use GPAI systems for consequential decision-making?

Only if they can be evaluated

… for the **specific** task they are being used for

… with a **clear understanding** of the scope of inputs

… in a way that is **reliable** and **reproducible**

and we can put accountability frameworks (explanations, recourse, and so on, in place)

So should we use GPAI systems for consequential decision-making?

Caveat: Research is ongoing.

Only if they can be evaluated

… for the **specific** task they are being used for

… with a **clear understanding** of the scope of inputs

… in a way that is **reliable** and **reproducible**

and we can put accountability frameworks (explanations, recourse, and so on, in place)