12.16.2025

AI in Government

& what happens *without* explainability

David Evan Harris
Chancellor’s Public Scholar, UC Berkeley

UC Berkeley Haas

AI Use cases in government

Two clear use cases emerge

Low-risk decisions

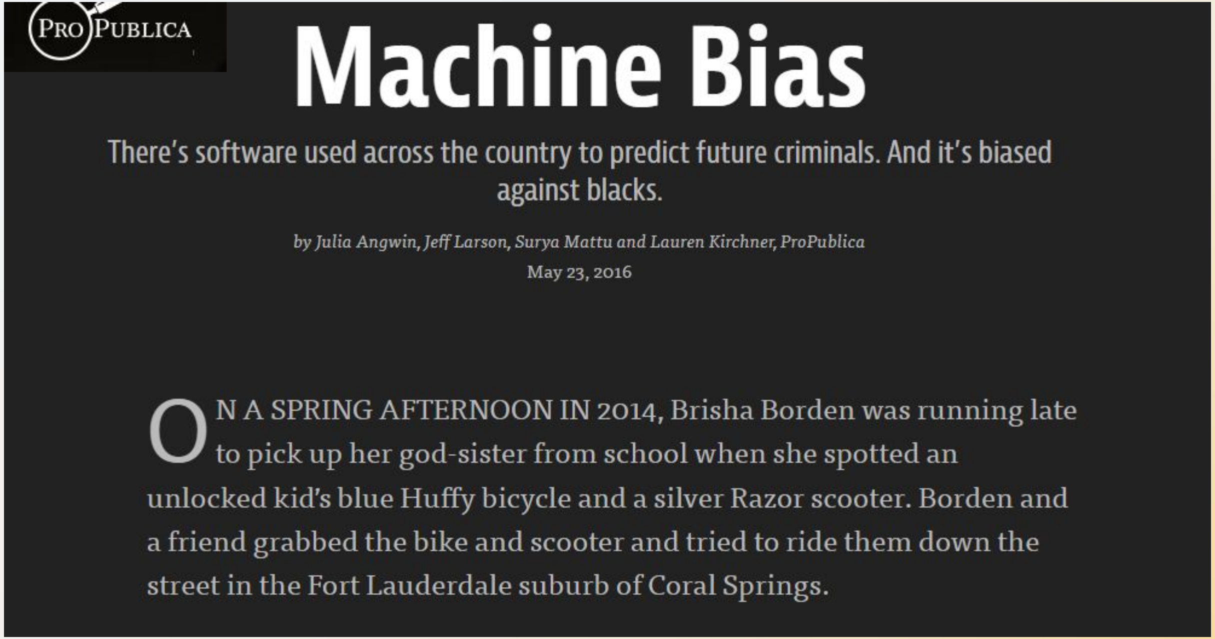
- Email & writing assistance without substantive research
- Grammar fixes
- Summarizing documents
- Database searches

High-risk decisions

- Employment decisions
- Housing and loan access
- Judicial decisions
- Policing
- Legal advice
- Decision-influencing analysis

Pre-2023 Automated Decision Systems (ADSs)

High-risk **ADS applications**



PROPUBLICA

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Angwin, Julia, Jeff Larson, Surya Mattu and Lauren Kirchner. "Machine Bias." *ProPublica*, May 23 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

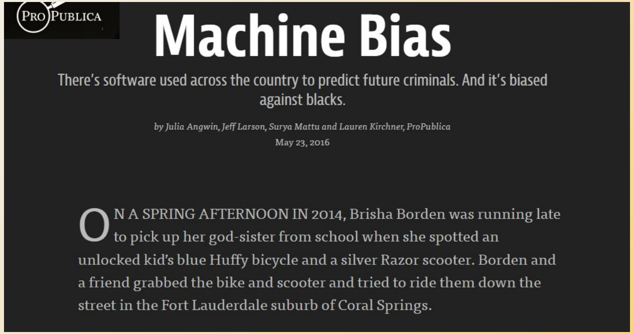
Pre-2023 Automated Decision Systems (ADSs)

High-risk **ADS applications**

*“We compared the recidivism risk categories predicted by the COMPAS tool to the actual recidivism rates of defendants in the two years after they were scored, and found that the score **correctly predicted an offender’s recidivism 61 percent** of the time, but was only **correct in its predictions of violent recidivism 20 percent** of the time.”*

Important to the investigation:

- **Defined** datasets
- **Reproducible** outcomes
- **Traceable** logic, algorithmic lines of reasoning



ProPublica

Machine Bias

There’s software used across the country to predict future criminals. And it’s biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid’s blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Angwin, Julia, Jeff Larson, Surya Mattu and Lauren Kirchner. “Machine Bias.” *ProPublica*, May 23 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

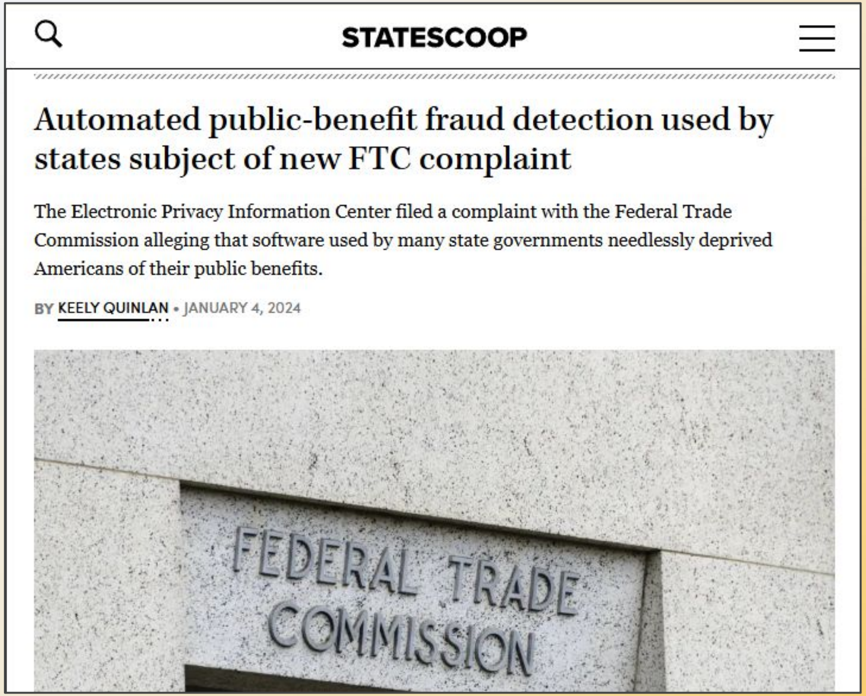
Pre-2023 Automated Decision Systems (ADSs)

High-risk **ADS applications**

SNAP in 42 states saw 1.1 million claimants' benefits suspended with ~**500,000 false positives**

Important to this ongoing case:

- **Defined** datasets
- **Reproducible** outcomes
- **Traceable** logic, algorithmic lines of reasoning



Automated public-benefit fraud detection used by states subject of new FTC complaint

The Electronic Privacy Information Center filed a complaint with the Federal Trade Commission alleging that software used by many state governments needlessly deprived Americans of their public benefits.

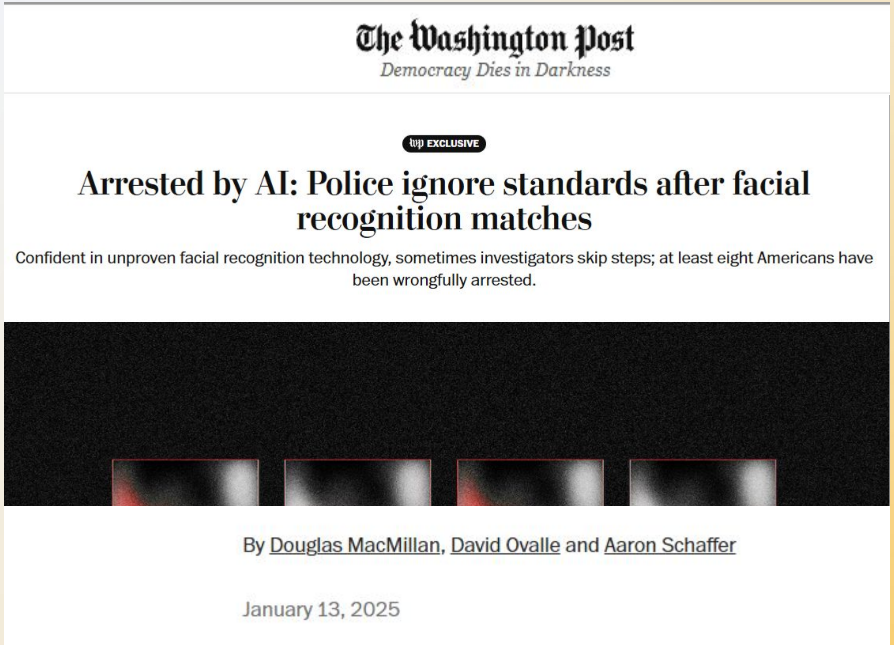
BY KEELY QUINLAN • JANUARY 4, 2024

Quinlan, Keely. "Automated public-benefit fraud detection used by states subject of new FTC complaint." *StateScoop*, Jan 4 2024. https://statescoop.com/automated-public-benefit-fraud-detection-state-ftc-complaint/

Pre-2023 Automated Decision Systems (ADSs)

High-risk **ADS applications**

*Arrested and jailed for a crime he says he didn't commit, it would take Gatlin more than two years to clear his name....**Gatlin is one of at least eight people wrongfully arrested in the United States after being identified through facial recognition.***



The Washington Post

Democracy Dies in Darkness

wp EXCLUSIVE

Arrested by AI: Police ignore standards after facial recognition matches

Confident in unproven facial recognition technology, sometimes investigators skip steps; at least eight Americans have been wrongfully arrested.

By Douglas MacMillan, David Ovalle and Aaron Schaffer

January 13, 2025

Important to this ongoing case:

- Policing policy that "the results of the technology are 'nonscientific' and 'should not be used as the sole basis for any decision.'"

MacMillan, Douglas, David Ovalle and Aaron Schaffer. "Arrested by AI: Police ignore standards after facial recognition matches." *The Washington Post,* Jan 13 2025. https://www.washingtonpost.com/business/interactive/2025/police-artificial-intelligence-facial-recognition/

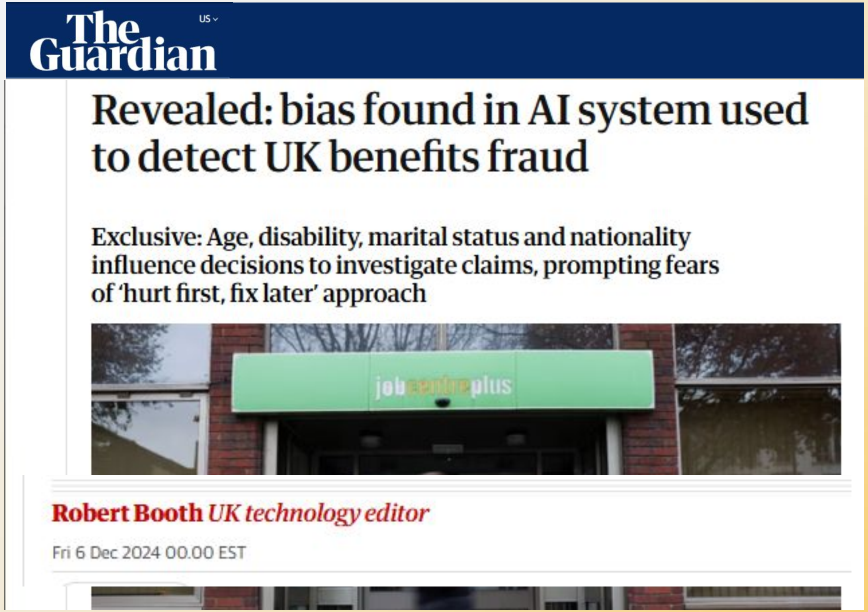
Pre-2023 Automated Decision Systems (ADSs)

High-risk **ADS applications**

"An internal assessment… found it incorrectly selected people from some groups more than others when recommending whom to investigate for possible fraud…. By one independent count, ***there are at least 55 automated tools being used by public authorities in the UK*** *potentially affecting decisions about millions of people"*

Important to this ongoing case:

- Defined datasets
- Reproducible outcomes
- Traceable logic, algorithmic lines of reasoning



The Guardian

US

Revealed: bias found in AI system used to detect UK benefits fraud

Exclusive: Age, disability, marital status and nationality influence decisions to investigate claims, prompting fears of 'hurt first, fix later' approach

Robert Booth *UK technology editor*

Fri 6 Dec 2024 00.00 EST

Booth, Robert. "Revealed: bias found in AI system used to detect UK benefits fraud." *The Guardian*, Dec 6 2024. https://www.theguardian.com/society/2024/dec/06/revealed-bias-found-in-ai-system-used-to-detect-uk-benefits

ADSs → General Purpose AI (GPAI)...

Challenge the core of explainability:

- Exceptionally large and ill-defined datasets
- Rarely reproducible results
- Rarely can provide logic or line of reasoning

Actively degrade explainability in logic:

- Generate narratives
- Post-hoc rationalizations

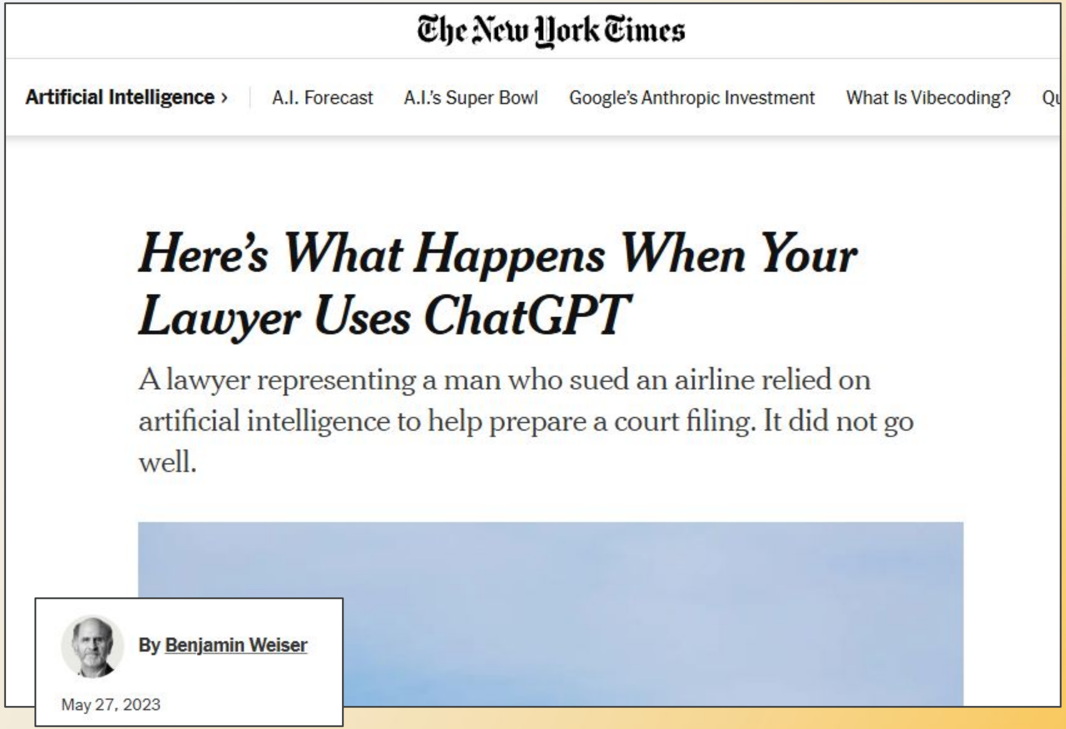
And overall, inhibits the accountability and correction process

High-risk **GPAI applications**

"Mr. Schwartz said that he had never used ChatGPT, and "therefore was unaware of the possibility that its content could be false." ***He had, he told Judge Castel, even asked the program to verify that the cases were real. It had said yes."***

What went wrong:

1. Logic failure→ hallucinations
2. Post-hoc rationalizations→ difficulty identifying errors
3. Poor use-case selection → dangerous decision-influencing analyses



The New York Times

Artificial Intelligence › | A.I. Forecast A.I.'s Super Bowl Google's Anthropic Investment What Is Vibecoding? Qu

Here's What Happens When Your Lawyer Uses ChatGPT

A lawyer representing a man who sued an airline relied on artificial intelligence to help prepare a court filing. It did not go well.

By Benjamin Weiser

May 27, 2023

Weiser, Benjamin. "Here's What Happens When Your Lawyer Uses ChatGPT." *New York Times*, May 27 2023. https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html

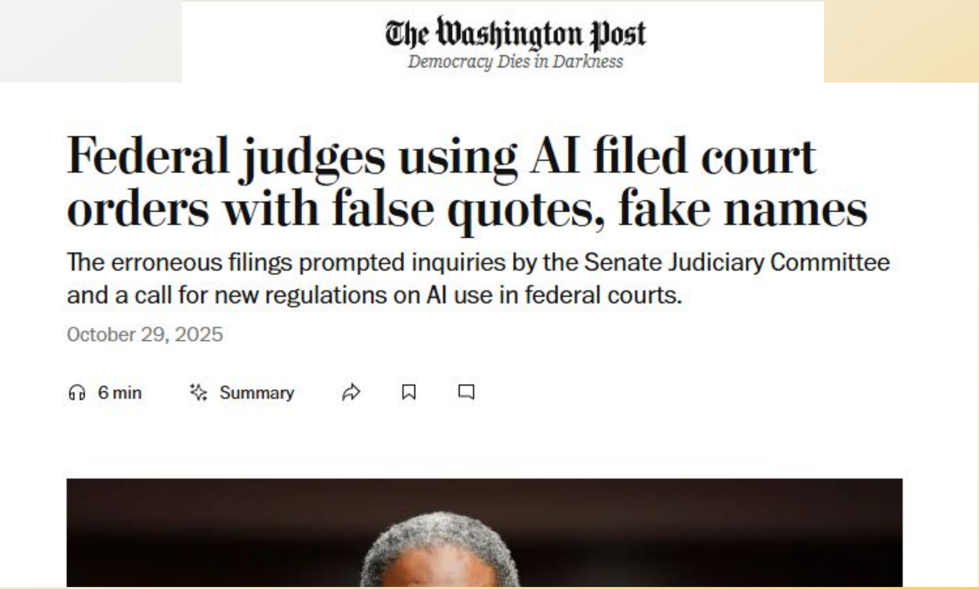
General Purpose Artificial Intelligence (GPAI)

High-risk **GPAI applications**

“Two federal judges in New Jersey and Mississippi admitted this month that their offices ***used artificial intelligence to draft factually inaccurate court documents that included fake quotes and fictional litigants*** *— drawing a rebuke from the head of the Senate Judiciary Committee.”*

What went wrong:

1. Logic failure→ hallucinations
2. Post-hoc rationalizations→ difficulty identifying errors
3. Poor use-case selection → dangerous decision-influencing analyses



The Washington Post
Democracy Dies in Darkness

Federal judges using AI filed court orders with false quotes, fake names

The erroneous filings prompted inquiries by the Senate Judiciary Committee and a call for new regulations on AI use in federal courts.

October 29, 2025

Wu, Daniel. “Federal judges using AI filed court orders with false quotes, fake names.” *The Washington Post*, Oct 29 2025. https://www.washingtonpost.com/nation/2025/10/29/federal-judges-ai-court-orders/

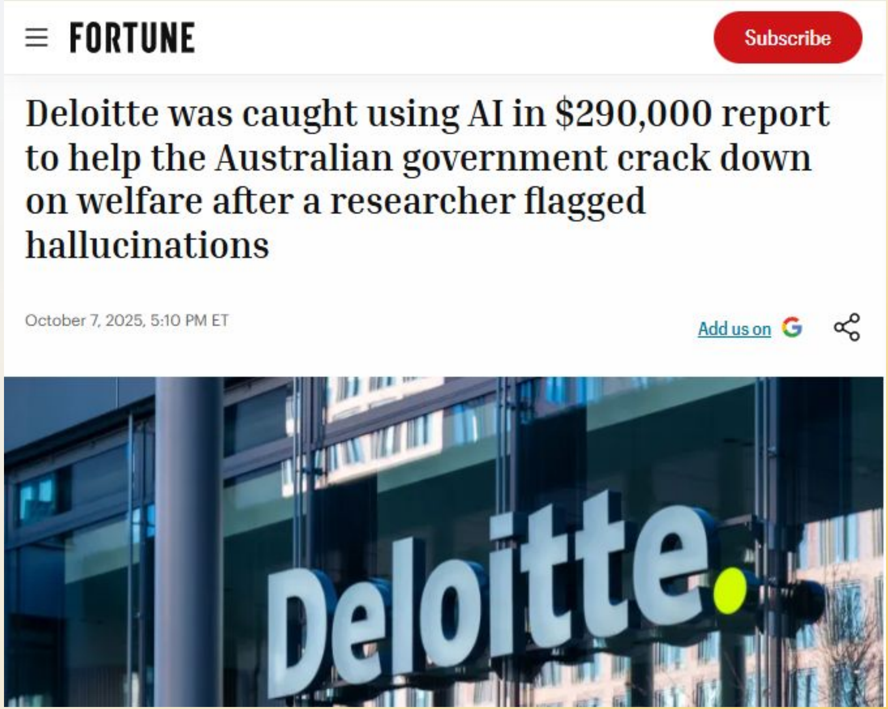
General Purpose Artificial Intelligence (GPAI)

High-risk **GPAI applications**

In audit report of Australian welfare system, **an LLM cited non-existent sources**

What went wrong:

1. Logic failure→ hallucinations
2. Post-hoc rationalizations→ difficulty identifying errors
3. Poor use-case selection → dangerous decision-influencing analyses



Deloitte was caught using AI in $290,000 report to help the Australian government crack down on welfare after a researcher flagged hallucinations

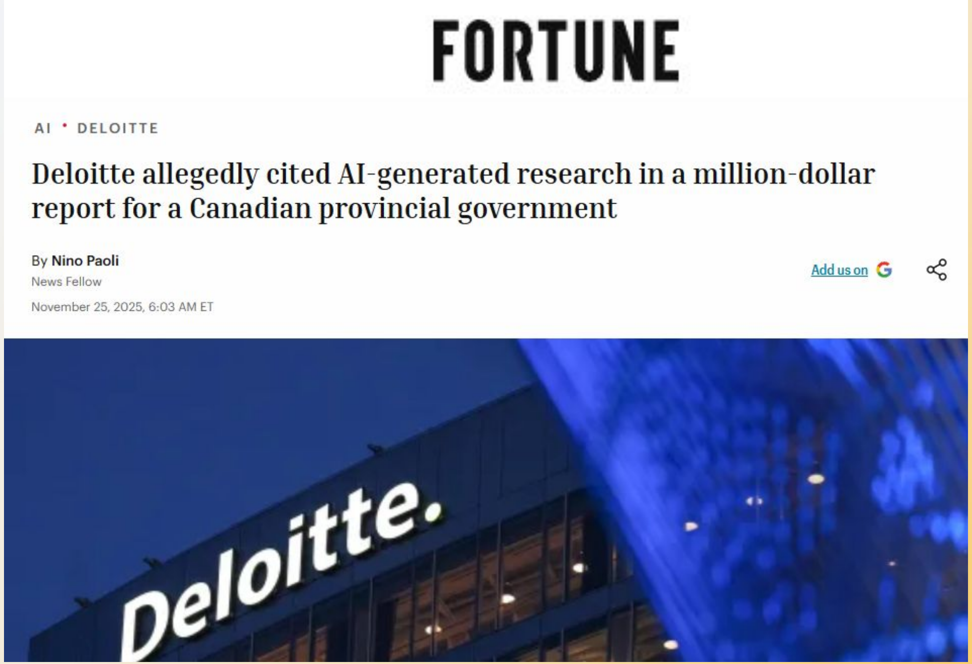
October 7, 2025, 5:10 PM ET

Paoli, Nino. “Deloitte was caught using AI in $290,000 report to help the Australian government crack down on welfare after a researcher flagged hallucinations.” *Fortune Magazine*, Oct 7 2025. https://fortune.com/2025/10/07/deloitte-ai-australia-government-report-hallucinations-technology-290000-refund/

General Purpose Artificial Intelligence (GPAI)

High-risk **GPAI applications**

“The Deloitte report ***contained false citations****, pulled from made-up academic papers* ***to draw conclusions for cost-effectiveness analyses,*** *and cited real researchers on papers they hadn’t worked on, the Independent found.”*



FORTUNE

AI • DELOITTE

Deloitte allegedly cited AI-generated research in a million-dollar report for a Canadian provincial government

By Nino Paoli
News Fellow
November 25, 2025, 6:03 AM ET

What went wrong:

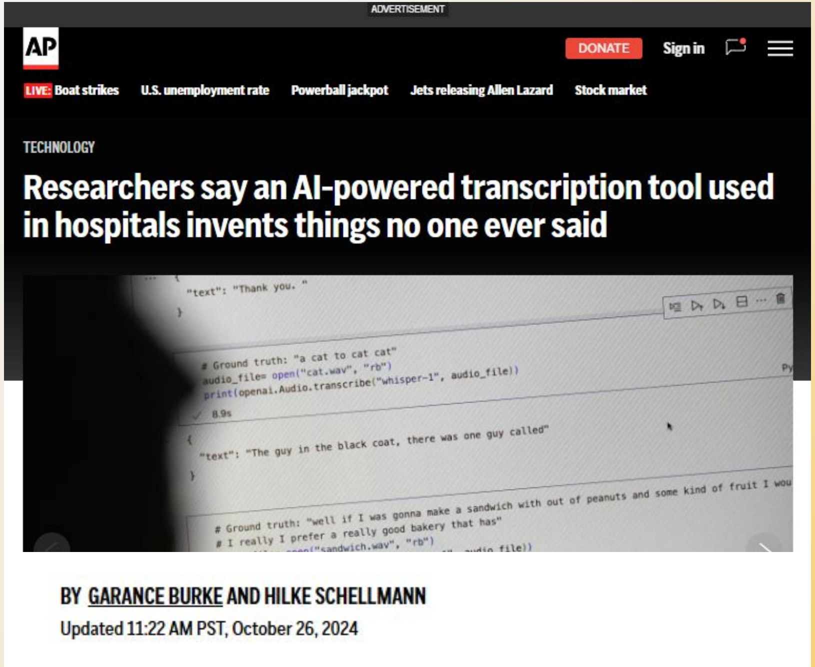
1. Logic failure→ hallucinations
2. Post-hoc rationalizations→ difficulty identifying errors
3. Poor use-case selection → dangerous decision-influencing analyses

Paoli, Nino. “Deloitte allegedly cited AI-generated research in a million-dollar report for a Canadian provincial government.” *Fortune Magazine,* Nov 25 2025. https://fortune.com/2025/11/25/deloitte-caught-fabricated-ai-generated-research-million-dollar-report-canada-government/

General Purpose Artificial Intelligence (GPAI)

High-risk **GPAI applications**

*"A machine learning engineer said he initially discovered **hallucinations in about half of the over 100 hours of Whisper transcriptions he analyzed**... More concerning, they said, is a rush by medical centers to utilize Whisper-based tools to transcribe patients' consultations with doctors, despite OpenAI' s warnings that the tool should not be used in 'high-risk domains.'"*



ADVERTISEMENT

AP

DONATE Sign in

LIVE: Boat strikes U.S. unemployment rate Powerball jackpot Jets releasing Allen Lazard Stock market

TECHNOLOGY

Researchers say an AI-powered transcription tool used in hospitals invents things no one ever said

```
    "text": "Thank you. "
}

# Ground truth: "a cat to cat cat"
audio_file= open("cat.wav", "rb")
print(openai.Audio.transcribe("whisper-1", audio_file))
✓ 8.9s
{
  "text": "The guy in the black coat, there was one guy called"
}

# Ground truth: "well if I was gonna make a sandwich with out of peanuts and some kind of fruit I wou
# I really I prefer a really good bakery that has"
```

BY GARANCE BURKE AND HILKE SCHELLMANN

Updated 11:22 AM PST, October 26, 2024

What went wrong:

1. Logic failure→ hallucinations
2. Post-hoc rationalizations→ difficulty identifying errors
3. Poor use-case selection → dangerous medical recordkeeping

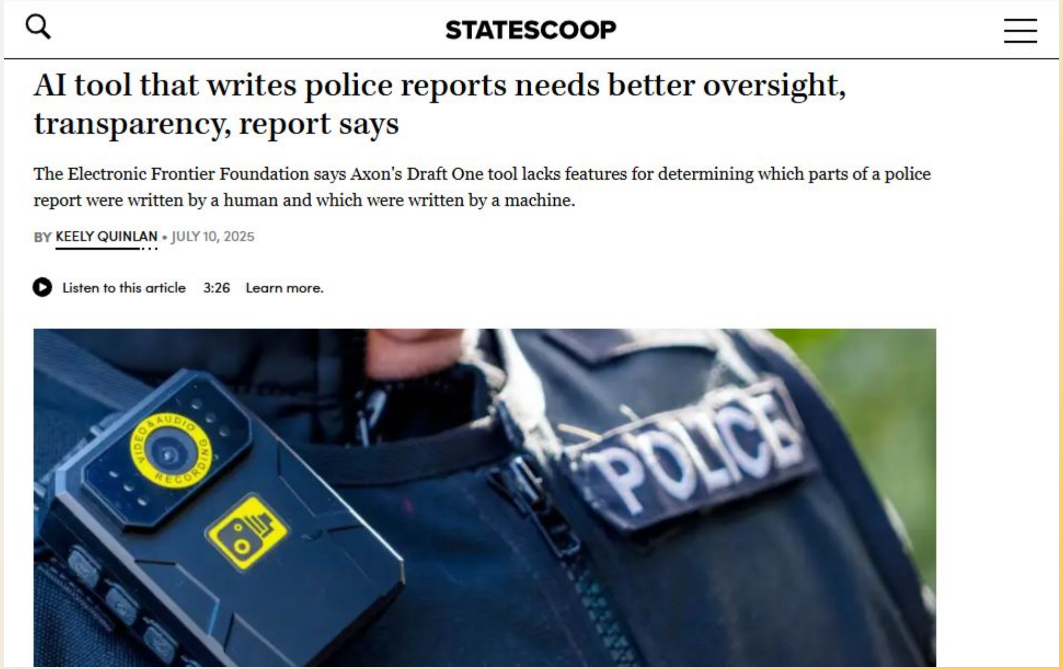
Burke, Garance, and Hilke Schellmann. "Researchers say AI transcription tool used in hospitals invents things no one ever said." *AP News*, Oct 26 2024. https://apnews.com/article/ai-artificial-intelligence-health-business-90020cdf5fa16c79ca2e5b6c4c9bbb14

High-risk **GPAI applications**

"the Palm Beach County Sheriff's Office, which requires a disclosure at the bottom of each police report if it was generated by AI, used Draft One to ***generate more than 3,000 reports*** *between December 2024 and March 2025...****Draft One's generative AI tech relies on a variation of OpenAI's ChatGPT to process body-worn camera audio,*** *and it creates police reports based only on the dialogue that is captured."*

What COULD go wrong:

1. Logic failure→ hallucinations
2. Post-hoc rationalizations→ difficulty identifying errors
3. Poor use-case selection → dangerous decision-influencing analyses


STATESCOOP

AI tool that writes police reports needs better oversight, transparency, report says

The Electronic Frontier Foundation says Axon's Draft One tool lacks features for determining which parts of a police report were written by a human and which were written by a machine.

BY KEELY QUINLAN • JULY 10, 2025

Listen to this article 3:26 Learn more.

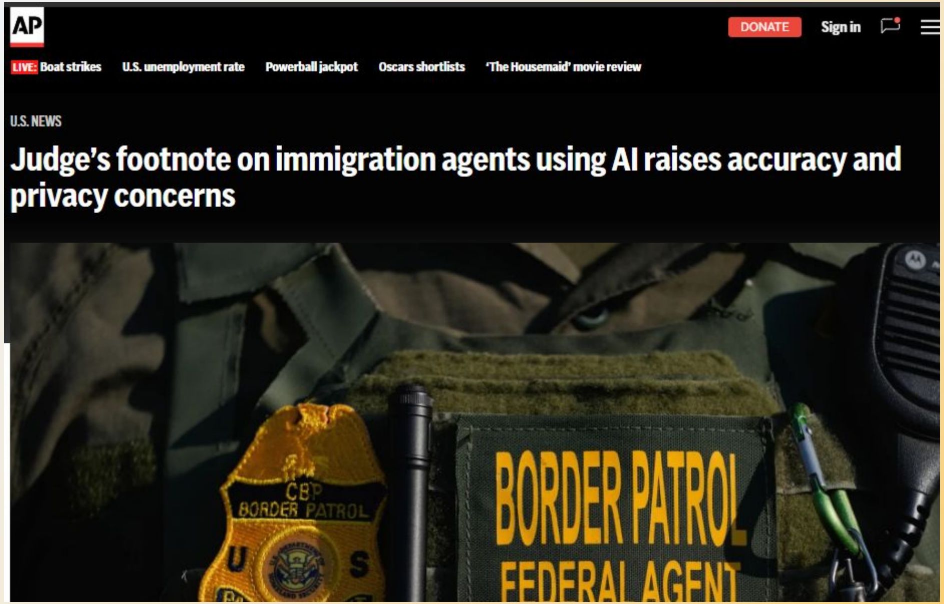
Quinlan, Keely. "AI tool that writes police reports needs better oversight, transparency, report says." *Statescoop,* July 10 2025. https://statescoop.com/gen-ai-police-report-transparency-oversight-eff/

High-risk **GPAI applications**

"[Judge Sara Ellis] Described what she saw in at least one body camera video, writing that an agent asks ChatGPT to compile a narrative… ***The judge noted factual discrepancies between the official narrative about those law enforcement responses and what body camera footage showed.****"*

What went wrong:

1. Logic failure→ hallucinations
2. Post-hoc rationalizations→ difficulty identifying errors
3. Poor use-case selection → dangerous decision-influencing analyses



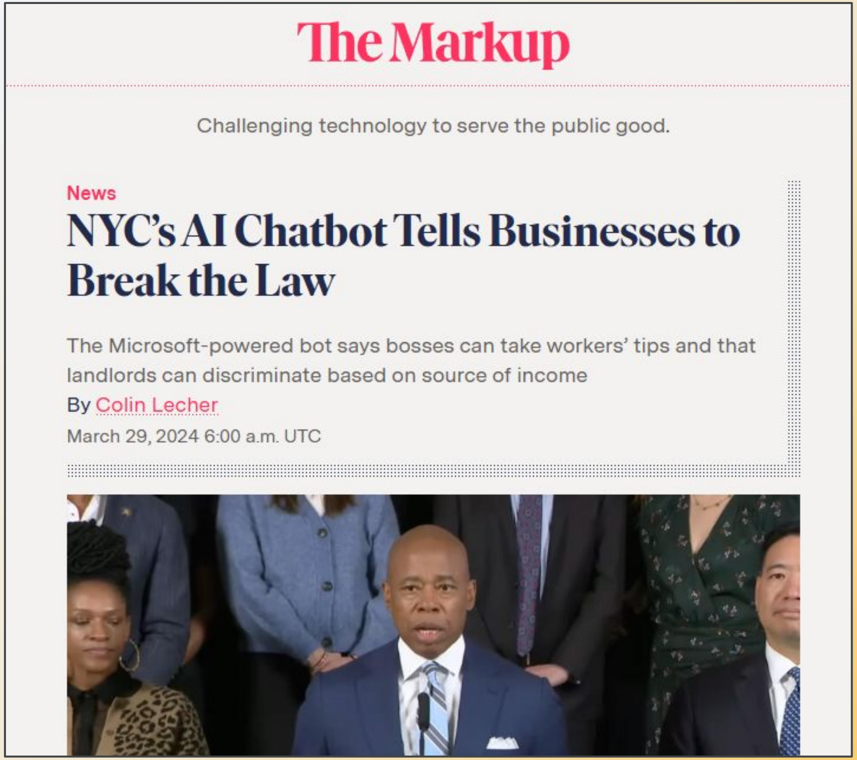
U.S. NEWS

Judge's footnote on immigration agents using AI raises accuracy and privacy concerns

Lauer, Claudia. "Judge's footnote on immigration agents using AI raises accuracy and privacy concerns." *AP News*, Nov 26 2025.
https://apnews.com/article/ice-artificial-intelligence-ai-chicago-law-enforcement-3b7aeb65c982842ce0b6b94436fbff30

High-risk **GPAI applications**

*"If you're a landlord wondering which tenants you have to accept, for example, you might pose a question like, "**are buildings required to accept section 8 vouchers?**" or "do I have to accept tenants on rental assistance?" In testing by The Markup, **the bot said no, landlords do not need to accept these tenants**. Except, in New York City, it's illegal for landlords to discriminate by source of income..."*



The Markup

Challenging technology to serve the public good.

News

NYC's AI Chatbot Tells Businesses to Break the Law

The Microsoft-powered bot says bosses can take workers' tips and that landlords can discriminate based on source of income

By Colin Lecher

March 29, 2024 6:00 a.m. UTC

What went wrong:

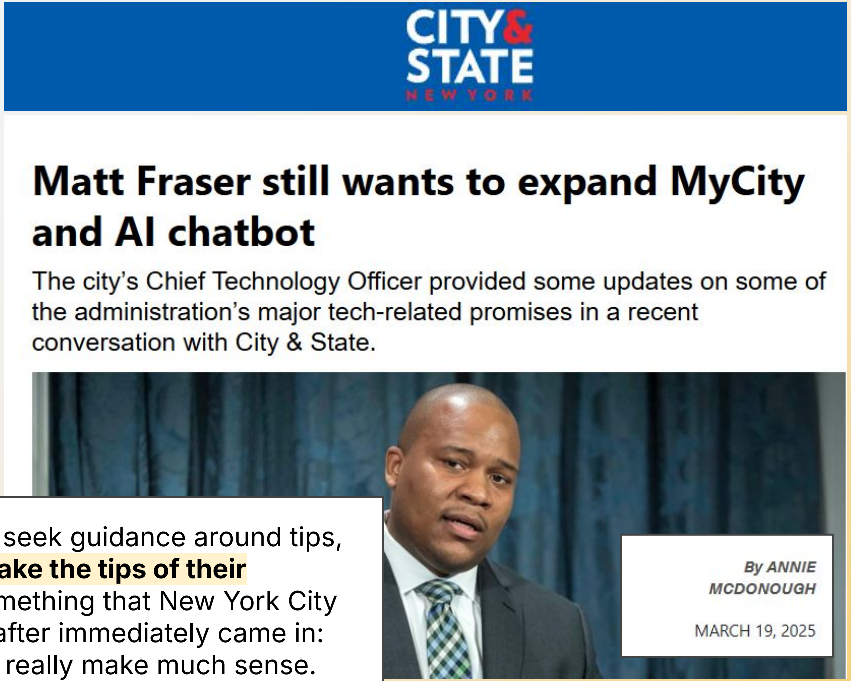
1. Logic failure→ hallucinations
2. Post-hoc rationalizations→ difficulty identifying errors
3. Poor use-case selection → dangerous legal advice

Lecher, Colin. "NYC's AI Chatbot Tells Businesses to Break the Law." *The Markup*, March 29 2024. https://themarkup.org/news/2024/03/29/nycs-ai-chatbot-tells-businesses-to-break-the-law

General Purpose Artificial Intelligence (GPAI)

Without explainability...

There is no accountability and correction process



Matt Fraser still wants to expand MyCity and AI chatbot

The city's Chief Technology Officer provided some updates on some of the administration's major tech-related promises in a recent conversation with City & State.

By ANNIE MCDONOUGH

MARCH 19, 2025

Instead of the accountability and correction process, **NYC simply lowered chatbot application from high-risk application to low-risk application**

"New York City's chatbot, when you seek guidance around tips, it says that **restaurant owners can take the tips of their employees**. Obviously, that's not something that New York City would push. And the requests thereafter immediately came in: take the chatbot down. That doesn't really make much sense. How about we fix it? So two weeks later, we put a patch in place, **and then we put a condition in place where anyone that asks anything outside the scope – redirect them and say it's not within the scope**" - Matt Fraser

McDonough, Annie. "Matt Fraser still wants to expand MyCity and AI chatbot." *City & State New York*, March 19 2025. https://www.cityandstateny.com/personality/2025/03/matt-fraser-still-wants-expand-mycity-and-ai-chatbot/403899/

What happens without accountability and correction processes?

Automated systems making consequential decisions incorrectly and with reduced accountability:

Societal consequences:

- Rapid erosion of trust in government processes
- Proliferation of societal bias and inequality

Physical consequences:

- Loss of access to food, housing, energy, water, health care

Legal consequences:

- Government violations of civil rights and other laws
- Litigation or class-action lawsuits
- Erosion of due process

High-risk **GPAI applications**

Takeaways:

1. Pre-2023 ADS already had many high-profile **accuracy and fairness failures**, but established **accountability and correction process** keep them in check and offer recourse

2. GPAIs suffer from **even worse accuracy issues** and have additional characteristics that obscure the information needed for accountability and correction process

3. As GPAIs are applied in situations where rules-based AI have already struggled, **we should expect more and more negative consequences**

4. Avoid this by **requiring use only of explainable systems for consequential decisions** and use **GPAI only in low-risk situations**

Thank you!

UC Berkeley Haas

Stay in Touch!

David Evan Harris

deh@berkeley.edu

LinkedIn.com/in/davidevanharris