AI Auditing

CIS 7000

Andrew Head & **Danaé Metaxa**

Recall

In the social media lecture, we described a bunch of examples of work identifying biases, e.g., impact on political news exposure

> E.g., studies measuring the diversity of news content on people's news feeds

But how do we measure these kinds of things in a robust, credible way?

Today

History of (Non-AI) Auditing

AI Auditing

AI Auditing Frontiers

History of Auditing

Old-skool, non-AI auditing

Auditing in the field

[Gaddis 2018]

Activists started conducting field experiments in the 1940s, visiting housing providers to see whether POC were treated equally.

Anti-discrimination law in 1960s in the US (Civil Rights Act of 1968) and UK (Race Relations Act) banned housing discrimination formally

Teams (white English, black English, immigrant) would apply for housing, employment, other services in person and see who succeeded

Most done by sociologists & economists

Second wave: scaling up

[Gaddis 2018]

Auditing continues through **correspondence studies**: sending applications by mail or fax rather than sending actors in person

Also include additional/more characteristics (sex, ability, age, education level, experience, skills)

Computerized resume creation facilitates this, and audits begin to become financially and logistically possible at larger scale

Expands to include political scientists and other social scientists

Are Emily and Greg more employable than Lakisha and Jamal?

[Bertrand & Mullainathan 2004]

Correspondence audit (mail-in applications)

Applied to 1300 job ads in Boston and Chicago, varying the name on the resume using birth record data to select names with racial connotations

Applicants with white-sounding names were 50% more likely to receive a callback than those with black-sounding names

Age, Women, and Hiring

[Lahey 2007]

Correspondence audit (mail-in applications)

One of the first to study age

Resumes for women ages 35 to 62 in MA and FL

A younger worker is 42-46% more likely to be offered an interview than an older one

Likely **statistical discrimination** (economic stereotyping) rather than **taste-based discrimination** (personal prejudice)

Third wave: online audits

[Gaddis 2018]

Early 2010s onward

Maintaining the same topics (employment, housing), the advent of **email and online application systems** moves auditing online

Domains also expand to include healthcare, politics (do constituents reaching out to legislators get replies?), education (how about students emailing professors?), and the new sharing economy (Airbnb, Uber)

Still mostly social scientists

Civ: Voter information

[White, Nathan, & Faller 2014]

Emails from Latino aliases are less likely to get voter information replies from election administrators.

Ed: University prestige

[Zeng & Luo 2025]

In China, students from prestigious universities get more, faster, more detailed, and friendlier replies than those at ordinary ones

Discrimination on Uber

[Ge et al., 2016]

More cancellations for white vs black-sounding passenger names, especially for men in low-density areas; longer, more expensive rides for women

…and on Airbnb

[Edelman, Luca, & Svirsky, 2017]

Applications to rent on Airbnb are 16% less likely to be accepted when the guest has a black-sounding name compared to a white-sounding one

Milestones in auditing

1960s: UK Parliament mandated oversight for anti-discrimination legislation

Milestones in auditing

1960s: UK Parliament mandated oversight for anti-discrimination legislation

1980s: Correspondence audits (e.g. Bertrand and Mullainathan, 2003)

Milestones in auditing

1960s: UK Parliament mandated oversight for anti-discrimination legislation

1980s: Correspondence audits (e.g. Bertrand and Mullainathan, 2003)

2000s: Automated correspondence audits (e.g. Oreopoulos, 2011)

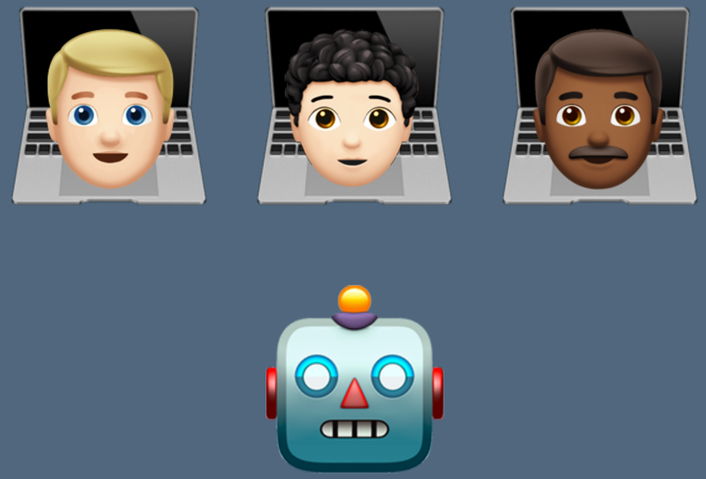
Milestones in auditing

1960s: UK Parliament mandated oversight for anti-discrimination legislation

1980s: Correspondence audits (e.g. Bertrand and Mullainathan, 2003)

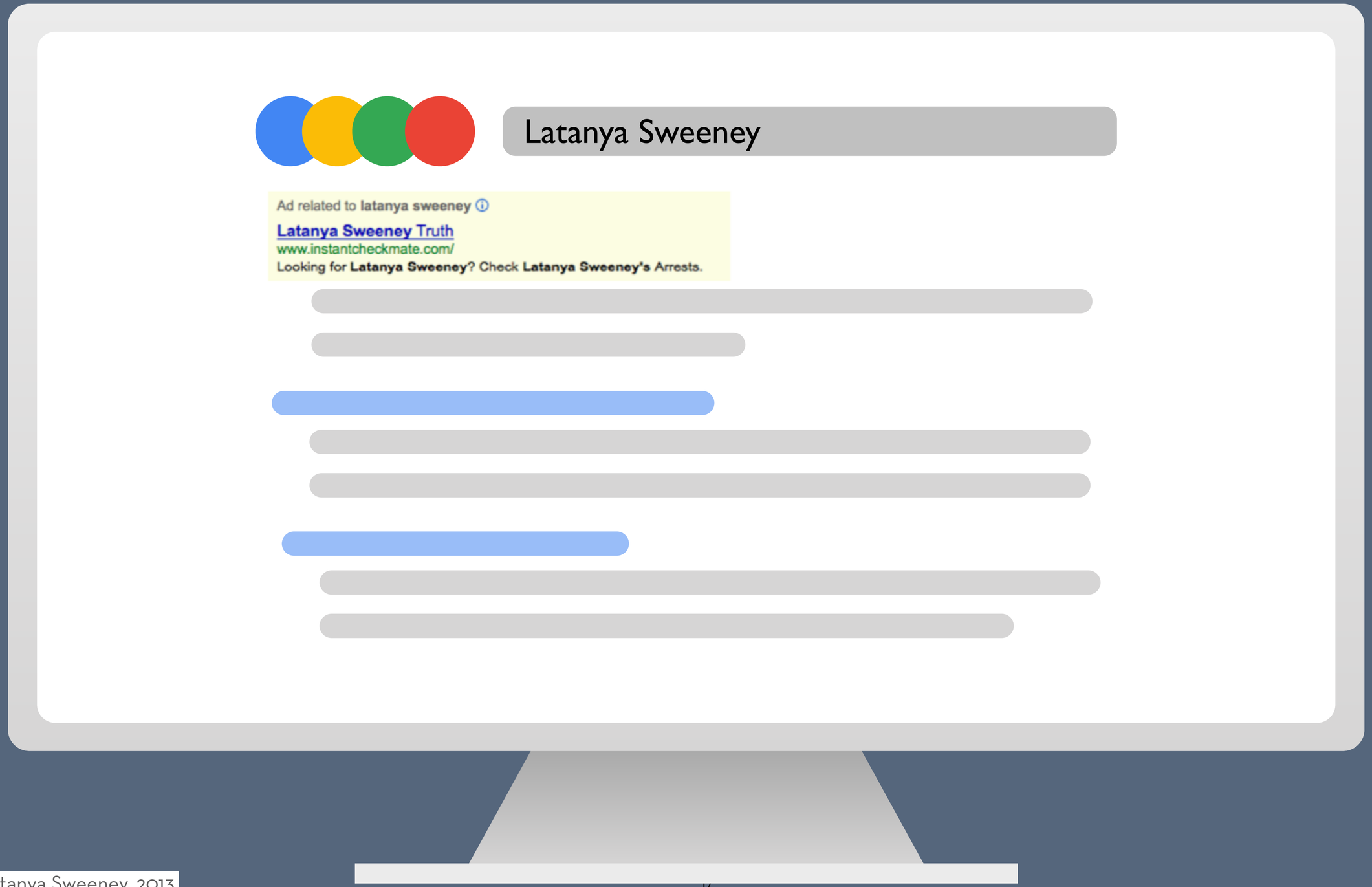
2000s: Automated correspondence audits (e.g. Oreopoulos, 2011)

…2010s: Auditing AI

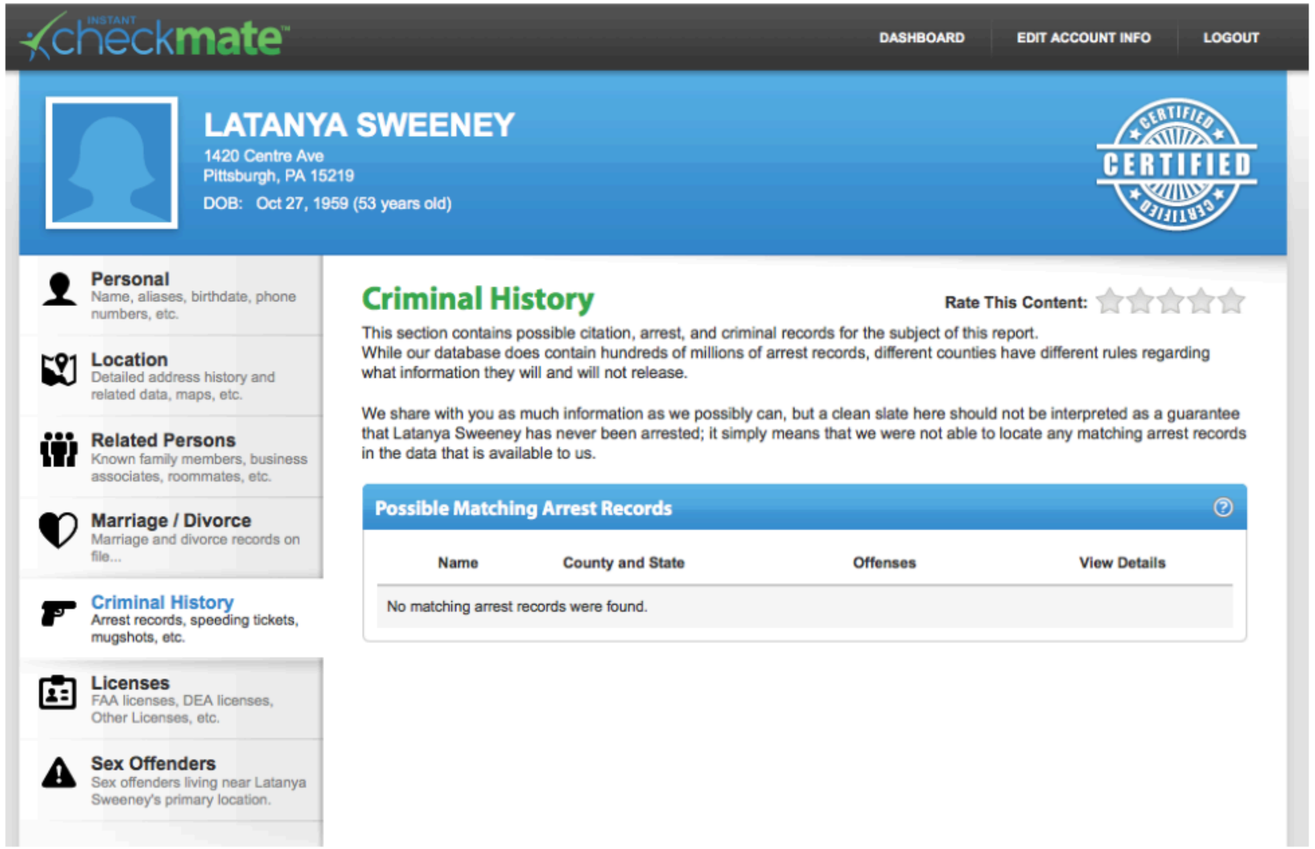


AI Auditing

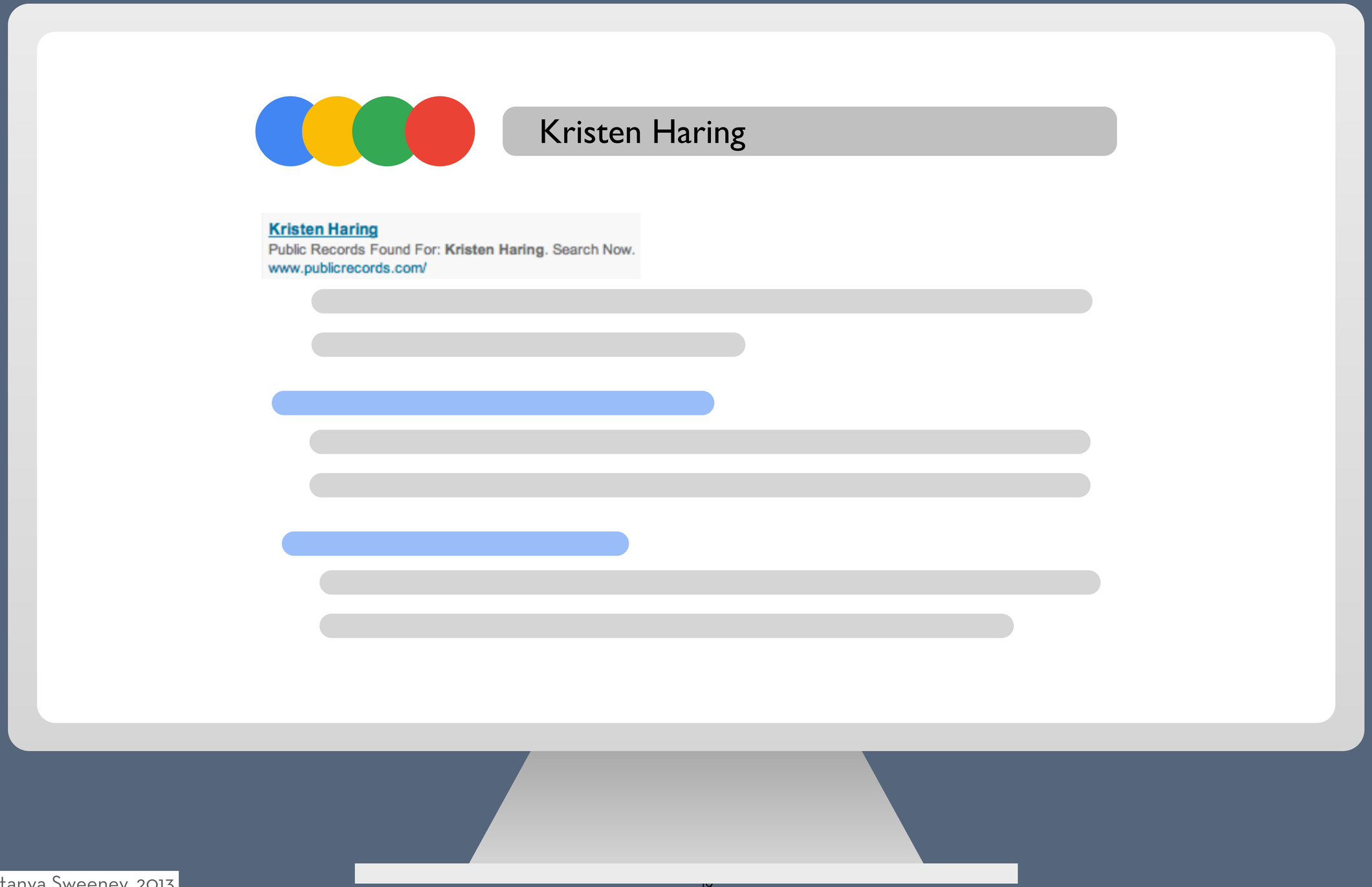
Beginning around 2010



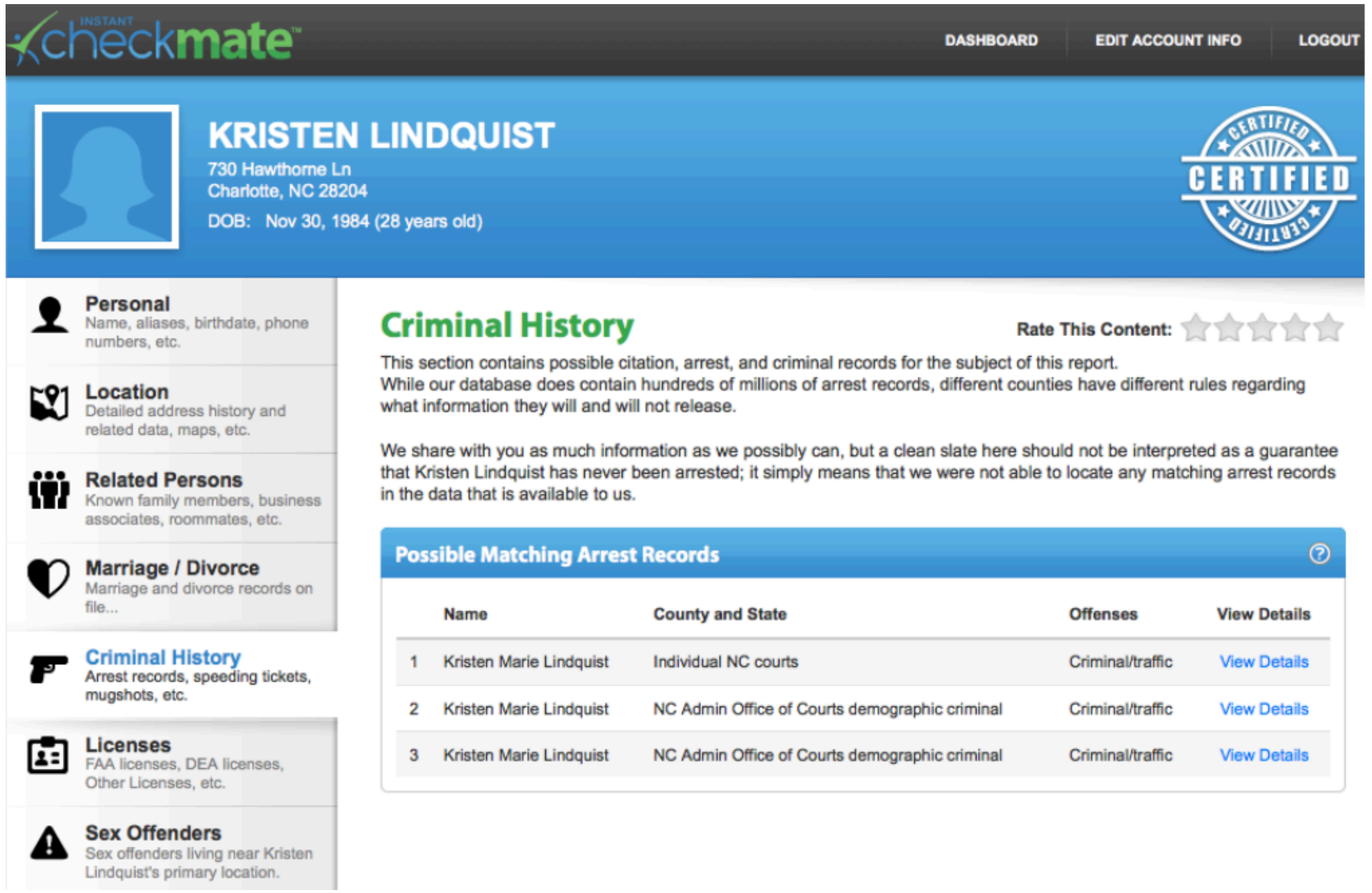
Images: Latanya Sweeney, 2013



Images: Latanya Sweeney, 2013



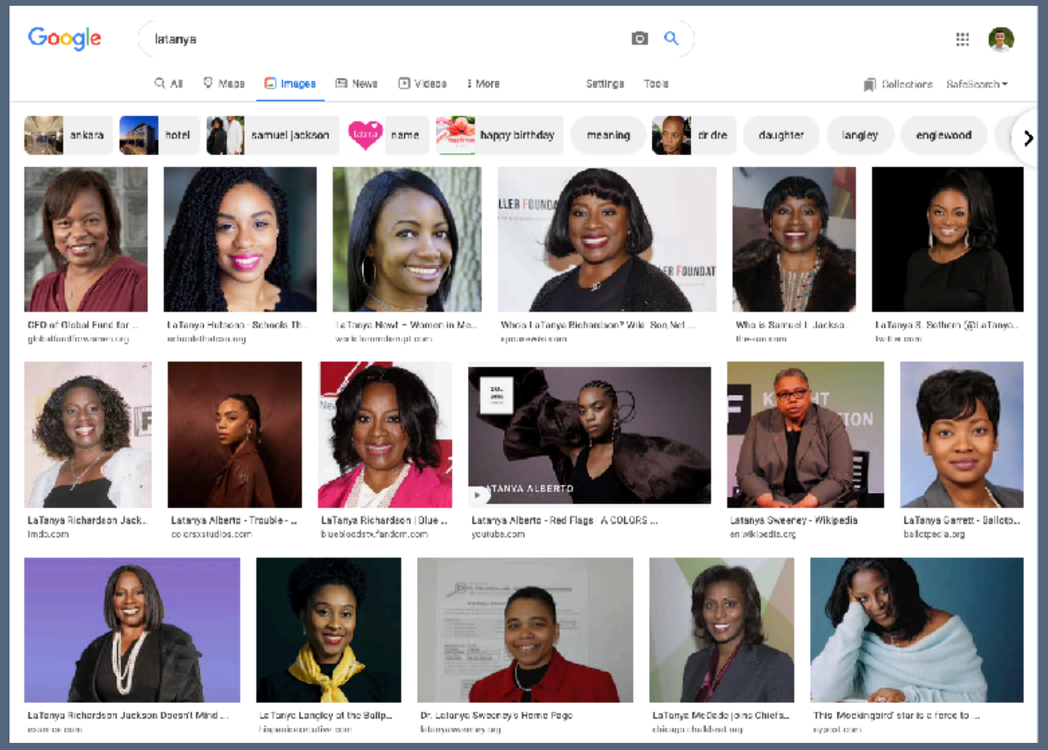
Images: Latanya Sweeney, 2013

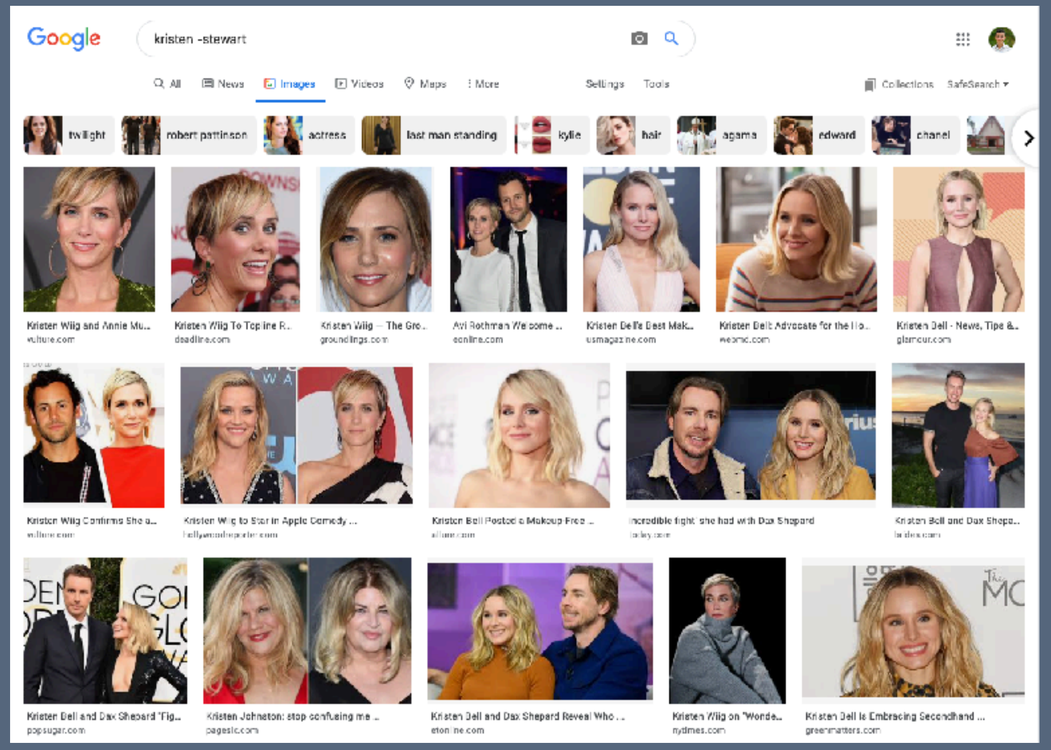


Images: Latanya Sweeney, 2013

The problem

Google Ads were 25% more likely to suggest an arrest record for Black-sounding names than white ones (Sweeney, 2013)





AI Auditing [Sandvig 2014]

YOU READ THIS

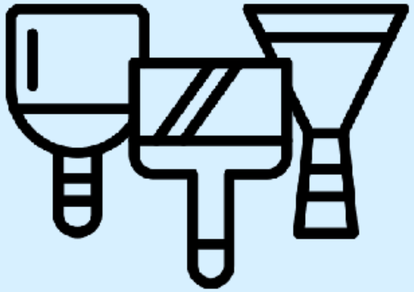
All the previous examples measure **humans discriminating against other humans**, even as methods and targets become increasingly computerized

As algorithmic systems began mediating access to housing, employment, information, etc: **can the audit be applied to AI?**

Sandvig describes 5 major audit types: code, noninvasive user, scraping, sock puppet, and crowdsourced auditing

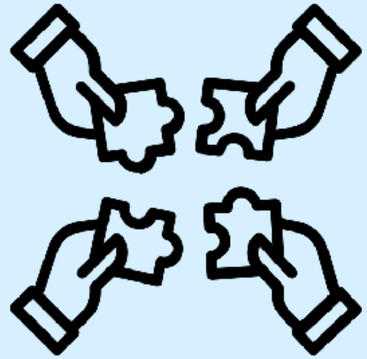
Goal: audit online platforms for discrimination

AI Auditing [Sandvig 2014]



Scraping audits

Sock puppets



Users involved

AI Auditing, expanded

[Metaxa et al. 2021]

What if the platform isn't discriminating, but is still nefarious? (e.g., popular rumor that Uber sets higher prices if your phone has low battery)

Updated Goal: draw conclusions about **how black-box systems work** using inference, without inside access

Evaluate systems w.r.t. implicit or explicit **standards** (social or legal)

Also, break down auditing process into key dimensions

Three components

The overall design of an algorithm audit has three main components:

1. Attribute–along what axis might the disparity exist?
2. Topic–what is the overall category or theme?
3. Platform–which algorithm or platform are we studying?

What should we audit?

What kinds of biases should we be auditing for in algorithms?

Race
Sex
Nationality
Religion
Disability

Gender
Sexuality
Political views
SES
Education

Employment
Health
Age
Pregnancy/
parent status

How should we select this attribute?

Legal ramifications, personal interest, ethical importance…

Common topics

Social science audits often focused on housing and employment (areas with **legal protections and concrete implications** if inequality was identified).

Common topics for AI audits have included:

- Housing
- Employment
- Gig economy
- Heathcare
- Consumer markets

Common platforms

The platform audited may very independently of the domain; domain is the topic, platform is where we're looking.

Common platforms include…

- Social media sites (FB, Twitter, etc.)
- Search engines (Google, Google Images, Bing, etc.)
- Commercial systems (health records, legal systems, etc.)
- Other online platforms (hiring sites, ads, etc.)

Other dimensions to consider

Once you've selected an axis of difference, topic, and algorithm, there are several other dimensions to keep in mind:

1. Temporal considerations (longitudinal? two points? single point?)
2. Data collection (manual, API, scraping, etc.)
   and annotation (Manual, crowdsourcing, ML, etc.)
3. Data analysis (paired t-test, fit model to ground truth, etc.)
4. Communicating findings (academic pub, general audience, direct)
5. Legal and ethical aspects (more later in a few slides!)

Notable examples

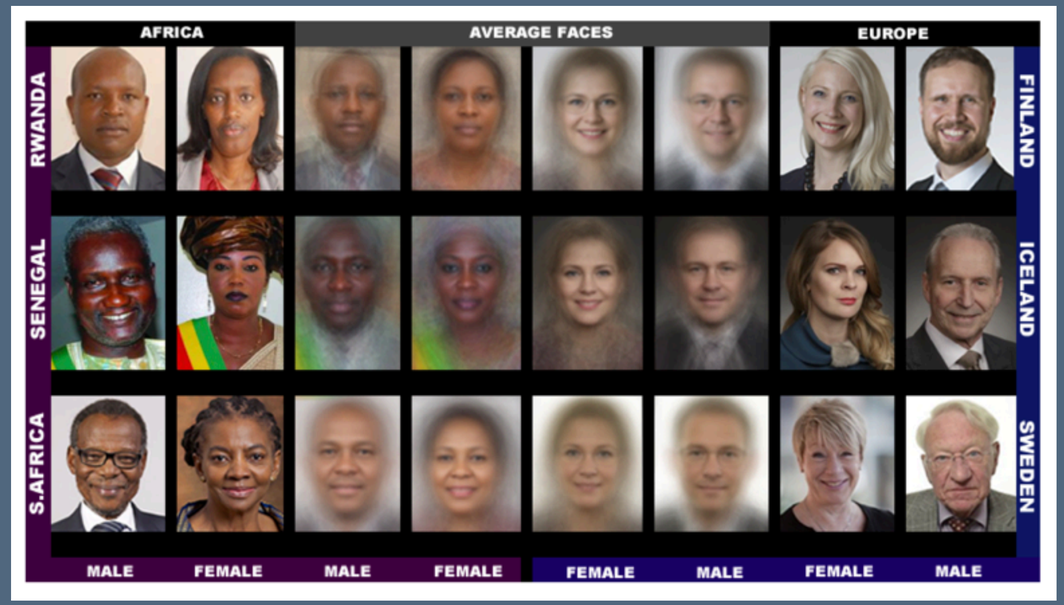
YOU READ THIS

Face recognition

[Buolamwini & Gebru, 2018]

Commercially-available face detection systems worked less well for darker-skinned, more feminine faces

Interesting followup [Raji et al 2020]: after contacting some companies, those services improved; those uncontacted did not (aka: they can fix it if they want to!)



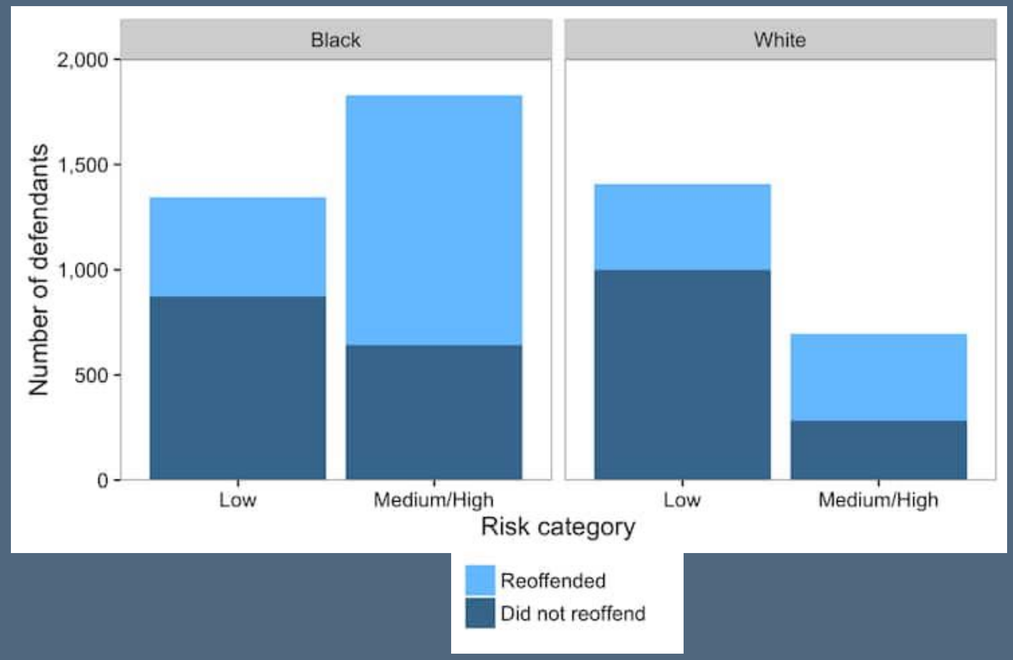
Criminal risk scores

[Angwin & Larson, 2016]

Analyzed data from a bail/sentencing algorithm to argue that risk scores were unequally assigned to defendants of different races

Identified disparate impact; some debate about whether that's the right metric, what the root problem is, etc.

Overall very high-profile example



Jobs & Ads

[Chen et al., 2018]

Collected resumes from hiring websites (recruiter perspective) and coded names for gender

Finding: slight penalty for female names (controlling for all else)

[Spiecher et al., 2018]

FB bans targeting by race/gender when advertising housing, jobs, etc

They show advertisers can easily still do discriminatory targeting via user attributes, free-form inputs, PII-based custom ads, and look-alike audiences

Healthcare

[Obermayer et al., Science 2019]

US health system uses tech to guide decisions, including a widely-used algo that assigns risk scores to patients

Using heath data, find that Black patients are sicker (more chronic illness) than white patients assigned the same score. **Why?**

Costs as proxy for needs!

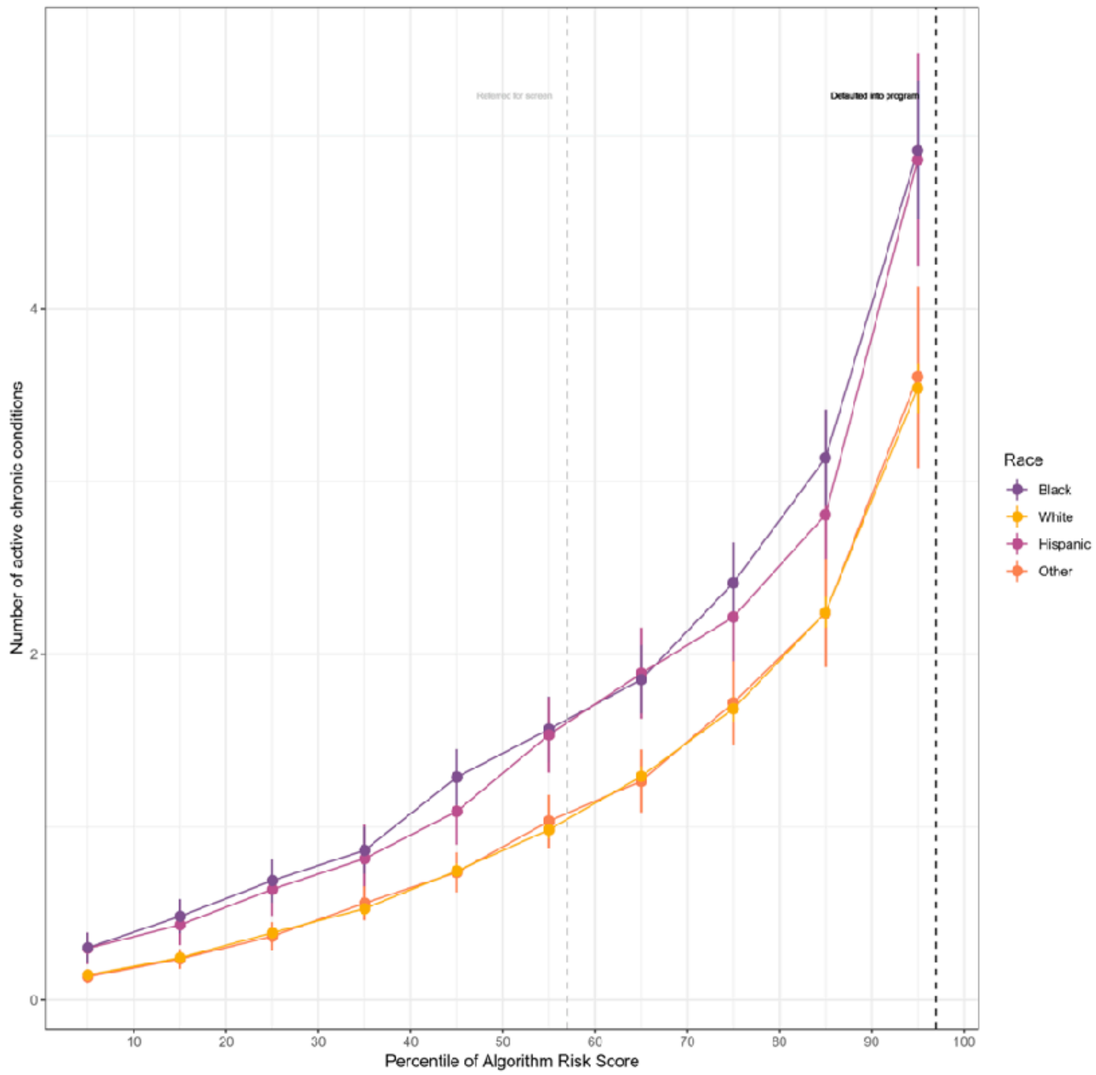


Fig. S1. Number of chronic illnesses vs. algorithm-predicted risk, including non-Black, non-White patients. Mean number of chronic conditions by race conditional on algorithm risk score.

Auditing LLMs

Several papers identifying race and gender biases when asking LLMs to make **hiring decisions** [Armstrong et al., 2024]

Audits of LLM **content moderation** layers find excessive censorship on identity-related topics and social issues [Proebsting et al., 2025; Dai et al., 2026]

LLMs fail in critical ways on complex policy topics, e.g., abortion legality [Encarnación & Metaxa, 2026]

…and so much more

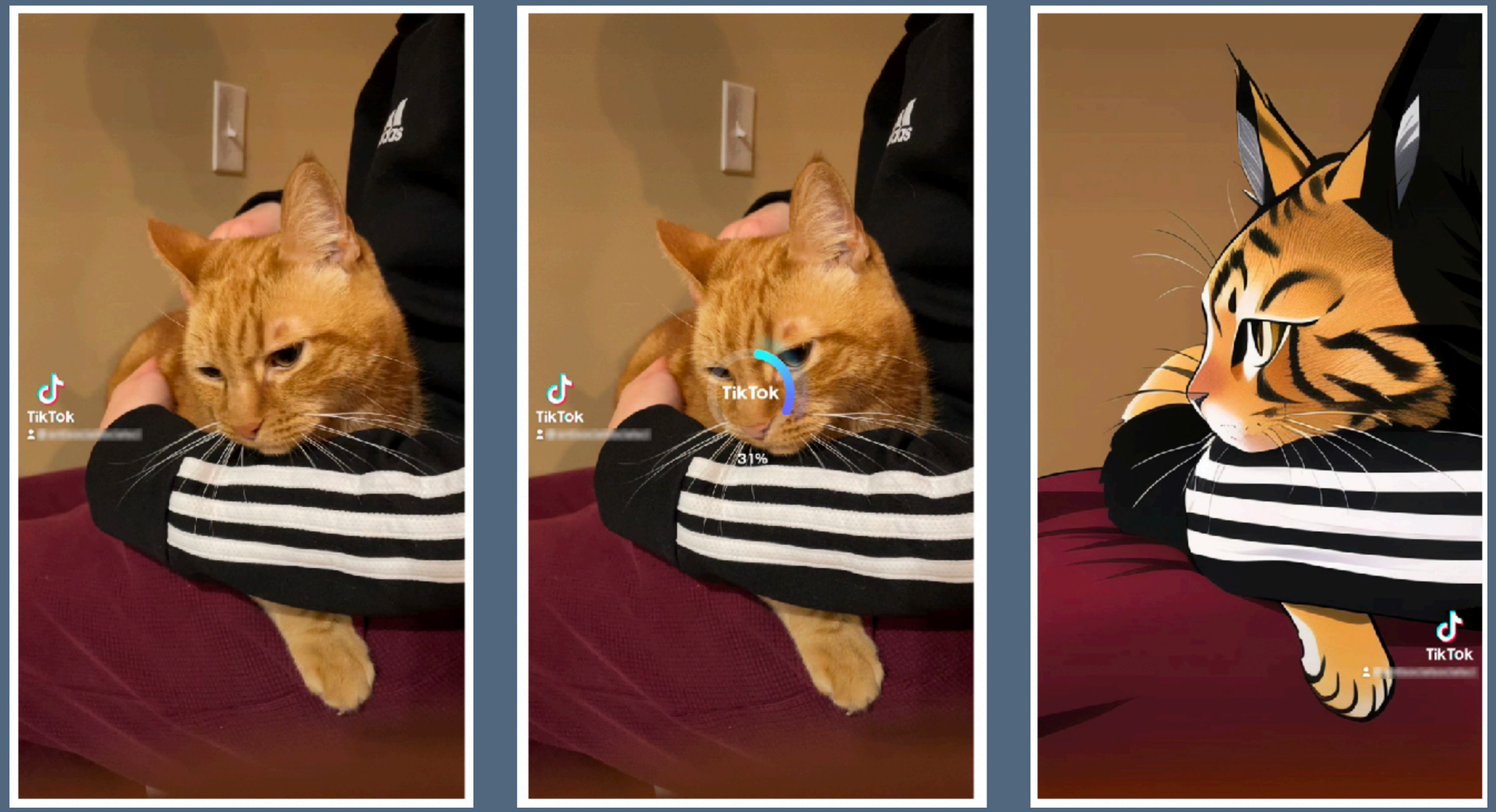
AI Auditing Frontiers

AI audits can be spurred by everyday life [Shen et al,. 2021]

"Everyday auditing [describes] cases in which **everyday users** of algorithmic systems detect and raise awareness about harmful behaviors that they encounter in the course of their **everyday interactions** with these systems."

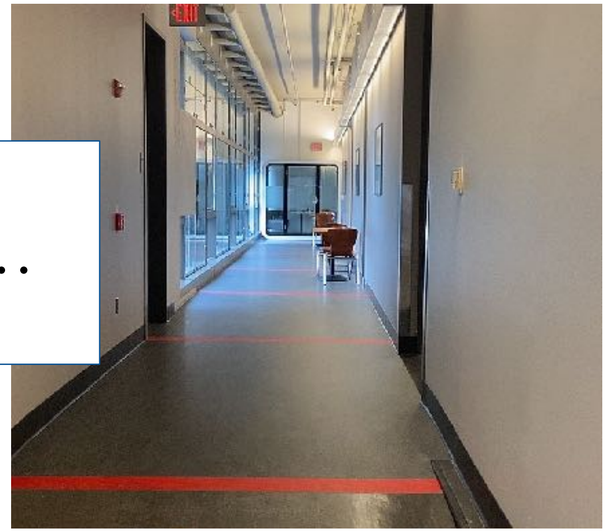
Ex: TikTok's AI Manga Filter

[Encarnación et al., 2026?]



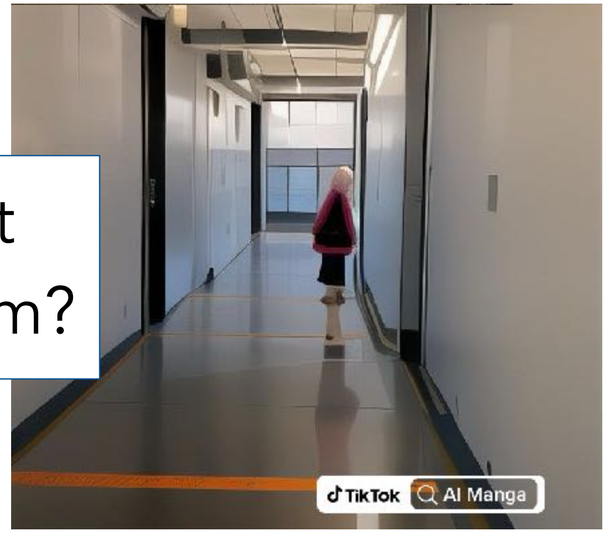
TikTok users noticing patterns…

Empty hallway…

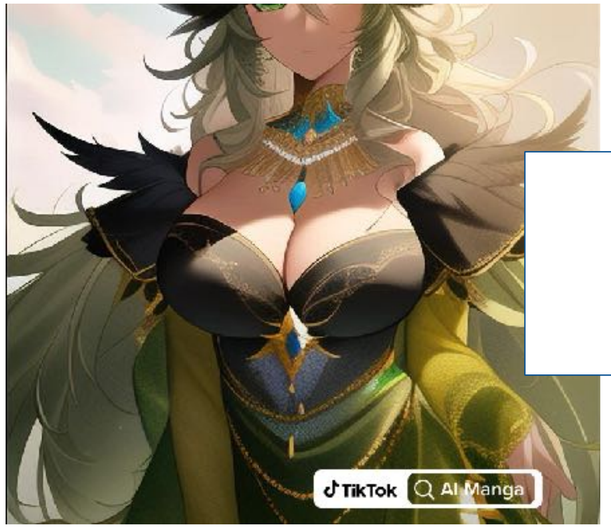


Ro’s face…

Where did that person come from?



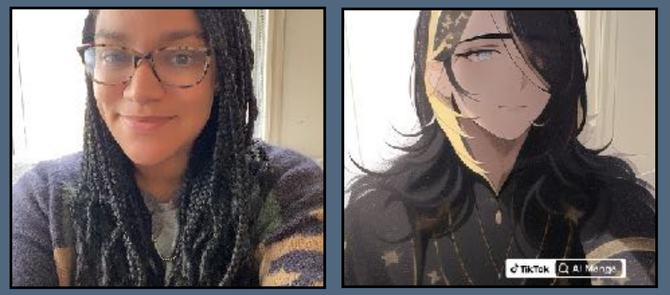
Definitely **not** Ro’s face.



Ex: TikTok's AI Manga Filter

[Encarnación et al., 2026?]

Anthropomorphization



Race & gender biases

Sexualization

Mandatory bias audits

In 2021, NYC Local Law 144 required employers using **automated employment decision tools** to publish bias audits of those tools

This was the first law in the US to mandate auditing! Other similar ones being considered in other jurisdictions

Sadly: it sucks. [Gerchick et al., 2025]

Very few actually published; many missing important demographic data or run on "test data" instead; some even published violations of the four-fifths rule; no enforcement.

The EU AI Act (2024)

Recent regulation in the EU; high-risk applications (employment, credit scoring, education, policing) must undergo assessments **before deployment**

Notable shift compared to the US where the most modern regulation still relies on external auditors post-deployment

Also gives citizens right to submit complaints and receive **explanations of AI decisions** that affect their rights

Problems there too, though

EU AI Act regulates open source AI less, but many providers are evading scrutiny by only being open weight, not open source [Liesenfeld & Dingemanse, 2024]

It also categorizes systems' risk according to "human vulnerability" but doesn't have a unified definition of what that means [Rebrean & Malgieri, 2025]

The Act safeguards health, safety, rights, democracy, law — but not socioeconomic status, among other attributes [Roy et al., 2025]

Auditing itself has limitations [Birhane et al., 2024]

Success: audits are no longer a niche academic thing; now done by regulators, law firms, civil society orgs, journalists, academics, and consulting firms

But getting from audit to impact is a challenge, and only a subset of audits translate into **accountability outcomes**

Summary

Audits are a method from social science used to uncover **discrimination in opaque human processes**, usually related to hiring or housing

Beginning in the 2010s, AI auditing emerged to apply the same method to **algorithmic decision-making** in the criminal system, healthcare, surveillance, and more have had significant impact

Auditing is beginning to be **required by law**, leading to a whole new era (with accompanying challenges) of auditing

Everyday users also engage in auditing behavior

References

ACM Proceedings. (n.d.). ACM Digital Library. Retrieved March 9, 2026, from https://dl.acm.org/proceedings

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May). Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bandy, J. (2021). Problematic machine behavior: A systematic literature review of algorithm audits. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1). https://doi.org/10.1145/3449148

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. American Economic Review, 94(4), 991–1013. https://doi.org/10.1257/000282804002561

Birhane, A., Steed, R., Ojewale, V., Vecchione, B., & Raji, I. D. (2024). AI auditing: The Broken Bus on the Road to AI Accountability. 2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), 612–643. https://doi.org/10.1109/SaTML59370.2024.00037

Buolamwini, J., & Gebru, T. (2018a). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Proceedings of Machine Learning Research, 81, 77–91.

Buolamwini, J., & Gebru, T. (2018b). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

References

Bushman, B. J., & Bonacci, A. M. (2004). You've got mail: Using e-mail to examine the effect of prejudiced attitudes on discrimination against Arabs. Journal of Experimental Social Psychology, 40(6), 753–759. https://doi.org/10.1016/j.jesp.2004.02.001

Butler, D. M., & Broockman, D. E. (2011). Do Politicians Racially Discriminate Against Constituents? A Field Experiment on State Legislators. American Journal of Political Science, 55(3), 463–477. https://doi.org/10.1111/j.1540-5907.2011.00515.x

Dai, Y., Lurie, E., Metaxa, D., & Friedler, S. A. (2025). Longitudinal Monitoring of LLM Content Moderation of Social Issues (arXiv:2510.01255). arXiv. https://doi.org/10.48550/arXiv.2510.01255

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. Digital Journalism, 3(3), 398–415. https://doi.org/10.1080/21670811.2014.976411

Edelman, B., Luca, M., & Svirsky, D. (2017). Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. American Economic Journal: Applied Economics, 9(2), 1–22. https://doi.org/10.1257/app.20160213

European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

Fix, M., & Struyk, R. J. (Eds.). (1993). Clear and Convincing Evidence: Measurement of Discrimination in America. Urban Institute Press.

References

Ge, Y., Knittel, C. R., MacKenzie, D., & Zoepf, S. (2016). Racial and Gender Discrimination in Transportation Network Companies (Working Paper No. 22776). National Bureau of Economic Research. https://doi.org/10.3386/w22776

Gerchick, M. K., Encarnación, R., Tanigawa-Lau, C., Armstrong, L., Gutiérrez, A., & Metaxa, D. (2025). Auditing the Audits: Lessons for Algorithmic Accountability from Local Law 144's Bias Audits. Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT '25, 29–44. https://doi.org/10.1145/3715275.3732004

Lahey, J. N. (2008). Age, Women, and Hiring: An Experimental Study. Journal of Human Resources, 43(1), 30–56. https://doi.org/10.3368/jhr.43.1.30

Liesenfeld, A., & Dingemanse, M. (2024). Rethinking open source generative AI: Open-washing and the EU AI Act. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24, 1774–1787. https://doi.org/10.1145/3630106.3659005

Metaxa, D., Gan, M. A., Goh, S., Hancock, J., & Landay, J. A. (2021). An image of society: Gender and racial representation and impact in image search results for occupations. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1). https://doi.org/10.1145/3449100

Metaxa, D., Park, J. S., Landay, J. A., & Hancock, J. T. (2019). Search media and elections: A longitudinal investigation of political search results in the 2018 U.S. elections. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW). https://doi.org/10.1145/3359231

References

Metaxa, D., Park, J. S., Robertson, R. E., Karahalios, K., Wilson, C., Hancock, J., & Sandvig, C. (2021). Auditing algorithms: Understanding algorithmic systems from the outside in. Foundations and Trends in Human-Computer Interaction, 14(4), 272–344. https://doi.org/10.1561/1100000083

New York City Council. (2021). New York City Local Law 144 of 2021: Automated Employment Decision Tools. https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

Pager, D. (2003). The mark of a criminal record. American Journal of Sociology, 108(5), 937–975. https://doi.org/10.1086/374403

Proebsting, G., Anigboro, O. I., Crawford, C. M., Metaxa, D., & Friedler, S. A. (2025). Identity-related Speech Suppression in Generative AI Content Moderation. Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO '25, 185–217. https://doi.org/10.1145/3757887.3763010

Raji, I. D., Xu, P., Honigsberg, C., & Ho, D. (2022). Outsider oversight: Designing a third party audit ecosystem for AI governance. Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 557–571. https://doi.org/10.1145/3514094.3534181

References

Rebrean, M.-L., & Malgieri, G. (2025). Vulnerability in the EU AI Act: Building an interpretation. Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT '25, 1985–1997. https://doi.org/10.1145/3715275.3732133

Roy, A., Rizou, S., Papadopoulos, S., & Ntoutsi, E. (2025). Achieving Socio-Economic Parity through the Lens of EU AI Act. Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT '25, 1890–1901. https://doi.org/10.1145/3715275.3732125

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and Discrimination: Converting Critical Concerns into Productive Inquiry, a Preconference at the 64th Annual Meeting of the International Communication Association.

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., & Langer, M. (2024). On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24, 2495–2507. https://doi.org/10.1145/3630106.3659051

Sweeney, L. (2013). Discrimination in online ad delivery. Communications of the ACM, 56(5), 44–54. https://doi.org/10.1145/2447976.2447990

Turner, M. A., Fix, M., & Struyk, R. J. (1991). Opportunities Denied, Opportunities Diminished: Racial Discrimination in Hiring. Urban Institute Press.

References

Vulnerability in the EU AI Act: Building an interpretation | Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. (n.d.). ACM Conferences. Retrieved March 9, 2026, from https://dl.acm.org/doi/10.1145/3715275.3732133

White, A. R., Nathan, N. L., & Faller, J. K. (2015). What Do I Need to Vote? Bureaucratic Discretion and Discrimination by Local Election Officials. American Political Science Review, 109(1), 129–142. https://doi.org/10.1017/S0003055414000562

Yinger, J. (1986). Measuring racial discrimination with fair housing audits: Caught in the act. American Economic Review, 76(5), 881–893.

Zeng, J., & Luo, X. (2025). Who gets a reply? An email-based field experiment on educational equality and university prestige. Higher Education. https://doi.org/10.1007/s10734-025-01572-3