Ethics, Public Policy, and Technological Change

Rob Reich
Mehran Sahami
Head TA: Roberta Fischli

Housekeeping

- Assignment #1 due tomorrow (Jan. 27) by 11:59pm.
  - Submit on Gradescope
- If you have OAE accommodations, please talk with your section leader about any extensions that may be needed on an assignment by assignment basis.

Today's Agenda

1. **Governance options**
2. Disclosure
3. Auditing
4. Alternatives

Fairness and Algorithms

Substantive fairness is context-dependent. It depends on some *social understanding* of what fairness requires.

- That's a hard problem to solve. It isn't a technical problem. It demands engagement in politics.

Procedural fairness is distinct from substantive fairness. We can envision what we might want a *process* to look like, especially from an "original position" (e.g. without knowing our own characteristics).

- This is what many policy efforts are focused on. Getting the details right is the challenge.

Remember the Policy Menu

Free market

Self-regulation
Private litigation
Disclosure
Licensing
Tax
Auditing
Premarket approval

Ban

ethics.
acm.org
ACM Code of Ethics and Professional Conduct

Remember the Policy Menu

Free market

Self-regulation
Private litigation
Disclosure
Licensing
Tax
Auditing
Premarket approval

Ban

New York City Council Proposal

James Vacca

Algorithmic Accountability bill:

- City agencies must publicly release source code of all algorithms for decisions
- Members of public can “self-test” algorithms (submit data, get results)

In Committee, bill watered down to:

- Establish task force to examine what algorithms are in use and whether they appear to discriminate
- Recommendations for public understanding and challenging results

Source: Craig Warga, New York Daily News, BetaNYC

A Glass Half Full? Or Half Empty?

Task force recommendations:

- Organizational structure and guidance for overseeing use of ADS
- Resources and support for agencies
- Broaden public discussion
- Formalize agency reporting on ADS, create process for assessing harm

Critics argue for more ambition:

- Public consultation before design / acquisition
- Standard and process for disparity assessments
- Right to sue (private action)
- Procedures for explanation of decisions (within 20 days)
- Transparency requirements

Imagine you are the Mayor

- Do you accept the recommendations? Why or why not?
- What would you prioritize to strengthen recommendations? Why?
- Why do you think these mechanisms are likely to be effective?

FAST COMPANY

The first effort to regulate AI was a spectacular failure

"City officials brought up the specter of unworkable regulations that would apply to every calculator and Excel document, a Kafkaesque nightmare where simply constructing a pivot table would require interagency approval. . . .

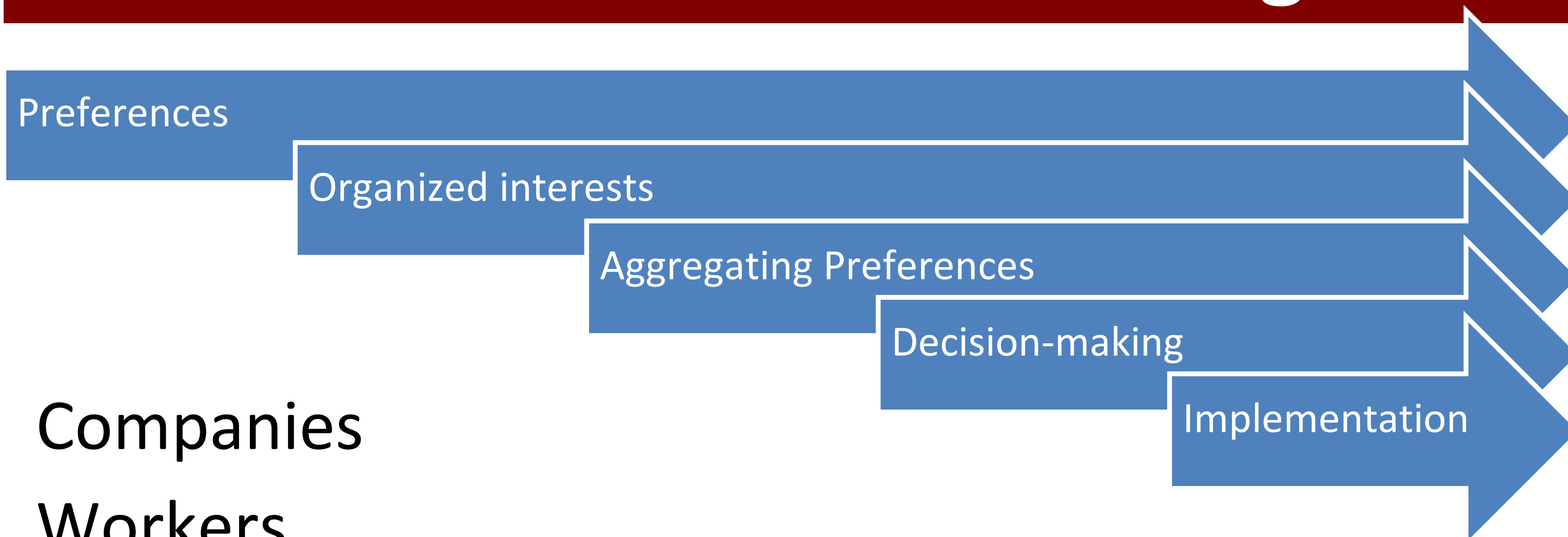
The [report] holds the air of a college paper hastily prepared by a student the day before the deadline.

Source: Cahn, Fast Company (2019)

Criteria

1. Efficiency: Does it improve public safety?
2. Fairness: Does it increase bias?
3. Transparency: Is it intelligible?
4. Due Process: Can people challenge its validity?
5. Privacy: Does it use private information?

Who Should be Governing?



Companies

Workers

Public

Consumers

Government

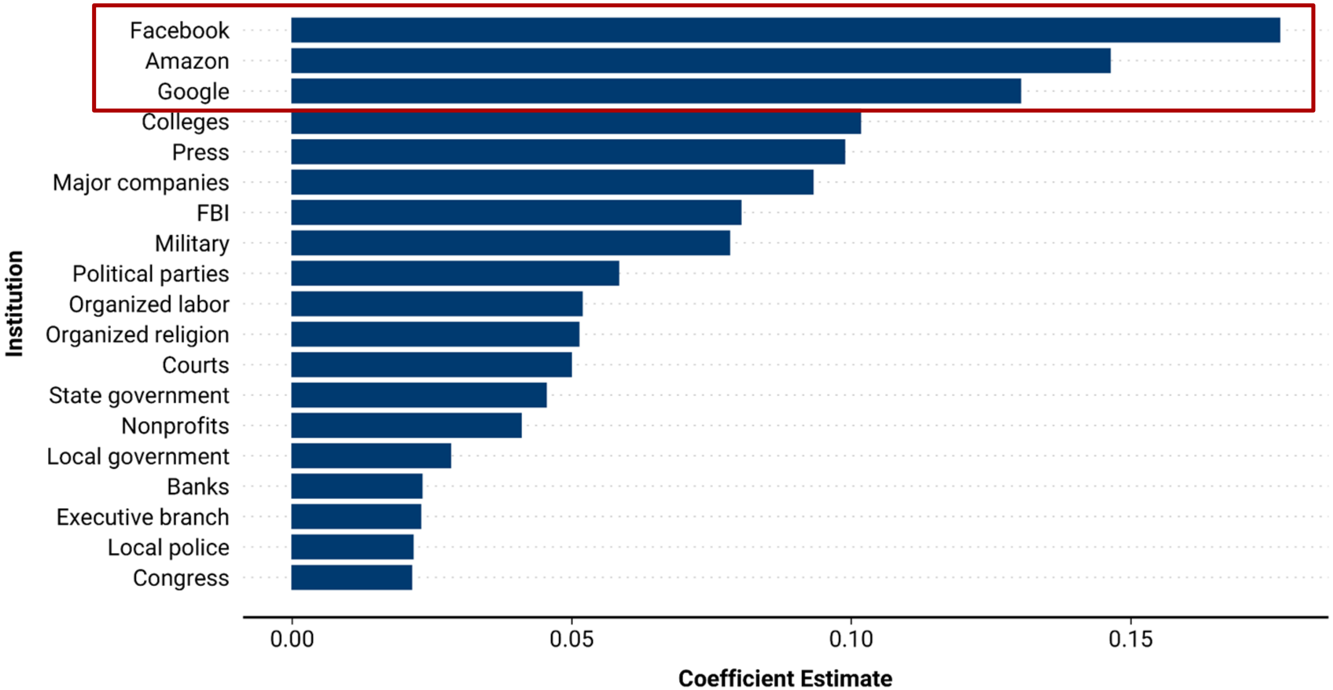
Courts

Companies

FIGURE 2

Loss in mean confidence between 2018 and 2021

We compare the sample-to-sample percentage loss in mean confidence of a series of US government, non-profit, and commercial institutions between the 2018 and 2021 American Institutional Confidence Polls.



Source: American Institutional Confidence Polls

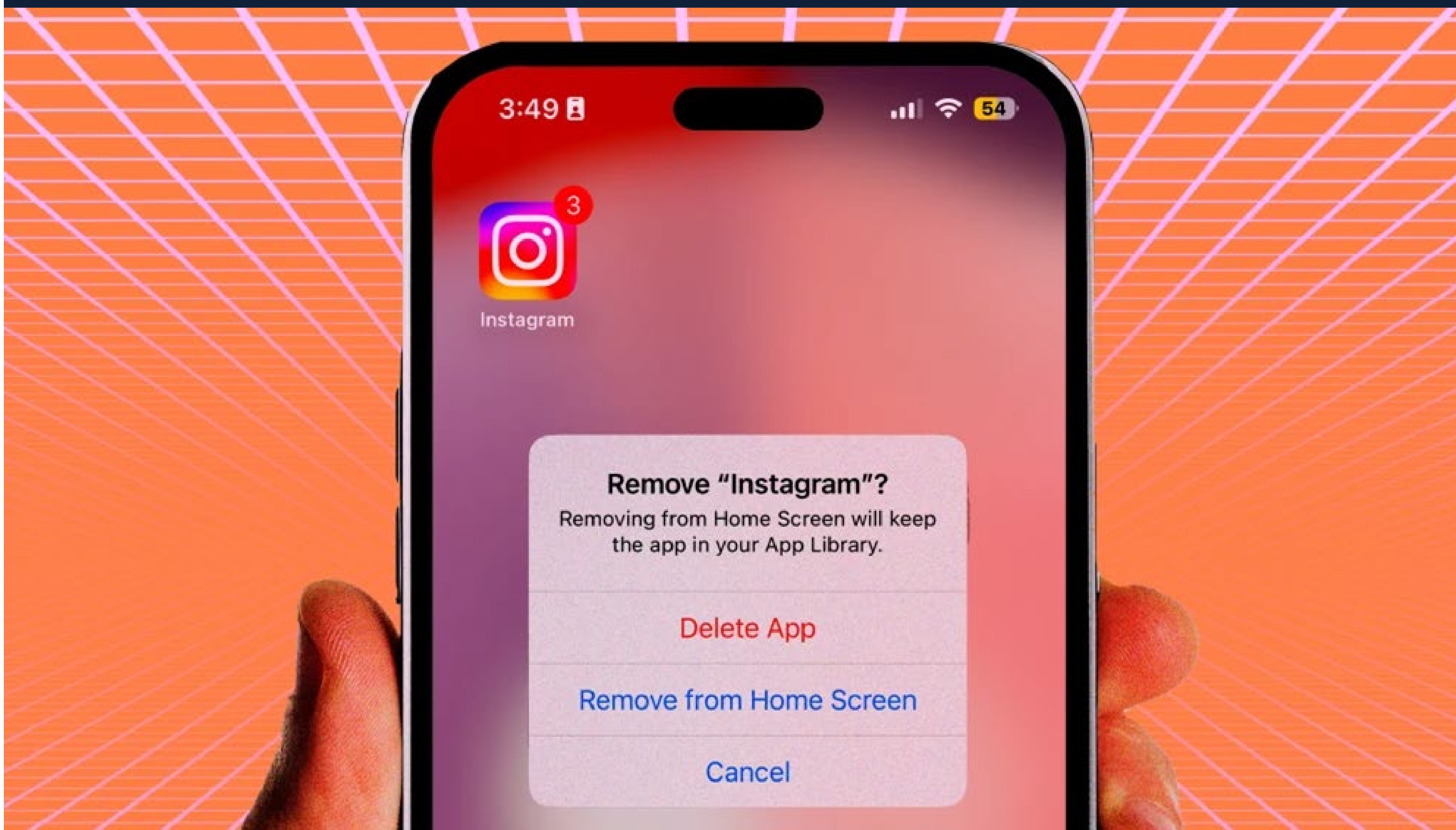
Source: Brookings

Workers

Source: New York Post

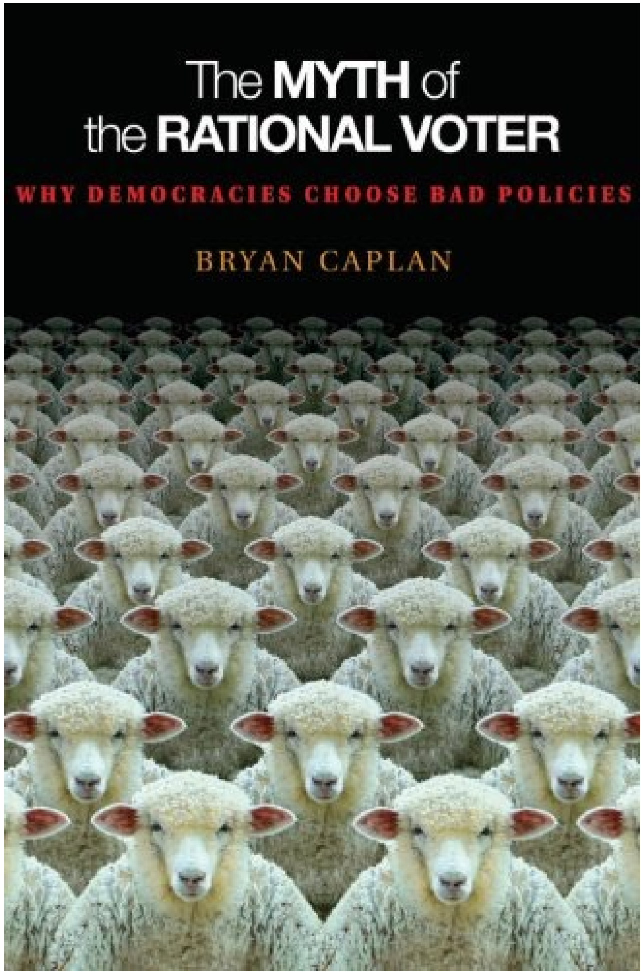
Users

Meta's pivot to the right sparks boycotts and calls for a user exodus



Source: NBC News

Public

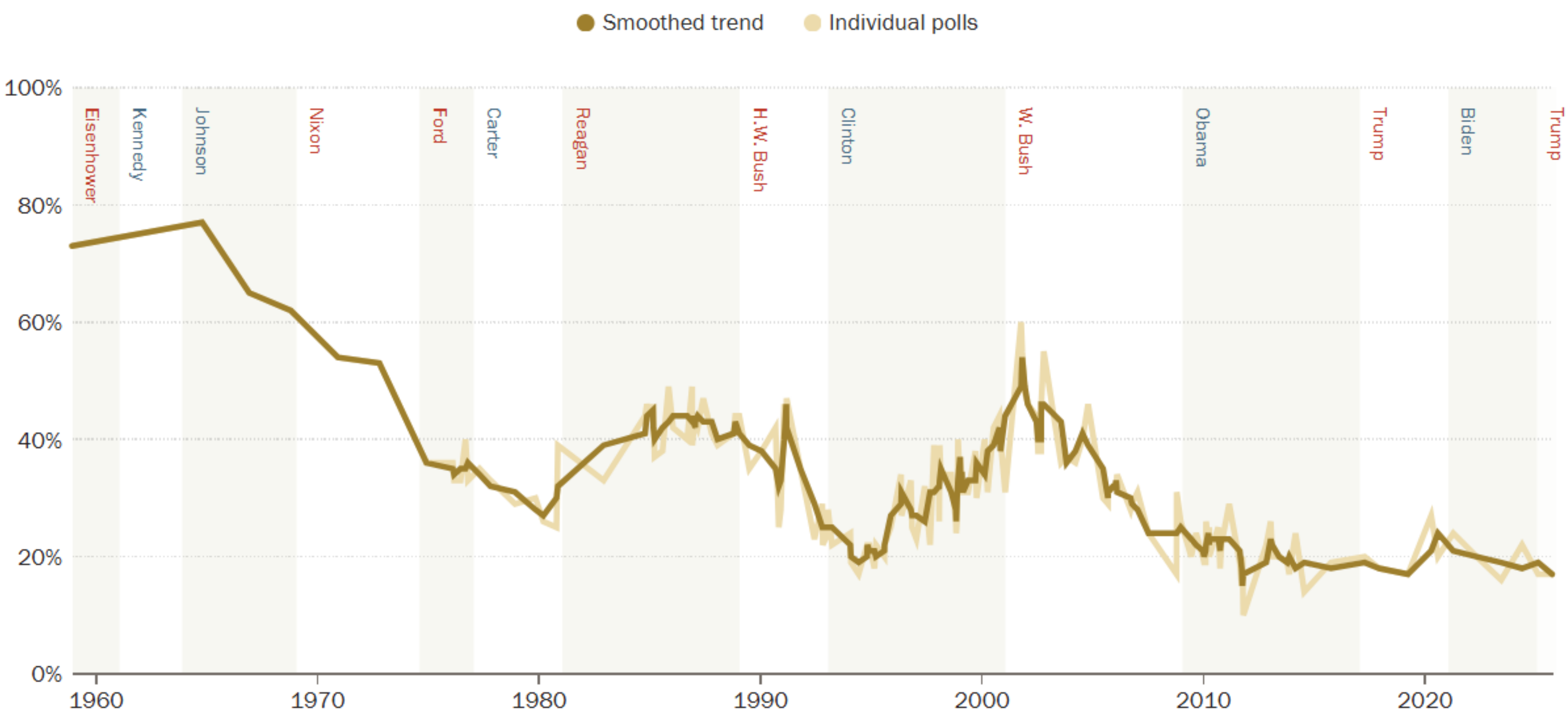


Sources: Bryan Caplan (2011), *The Myth of the Rational Voter*, Princeton University Press, Los Angeles Times

Government

Public trust in government near historic lows

% who say they trust the government in Washington to do what is right ***just about always/most of the time***



Note: From 1976-February 2025, the smoothed trend line represents a three-survey moving average. Data prior to 1976, and the most recent number (September 2025), are from individual polls.

Sources: Pew Research Center, National Election Studies, Gallup, ABC/Washington Post, CBS/New York Times, and CNN surveys.

PEW RESEARCH CENTER

Courts

Source: Vox

Today's Agenda

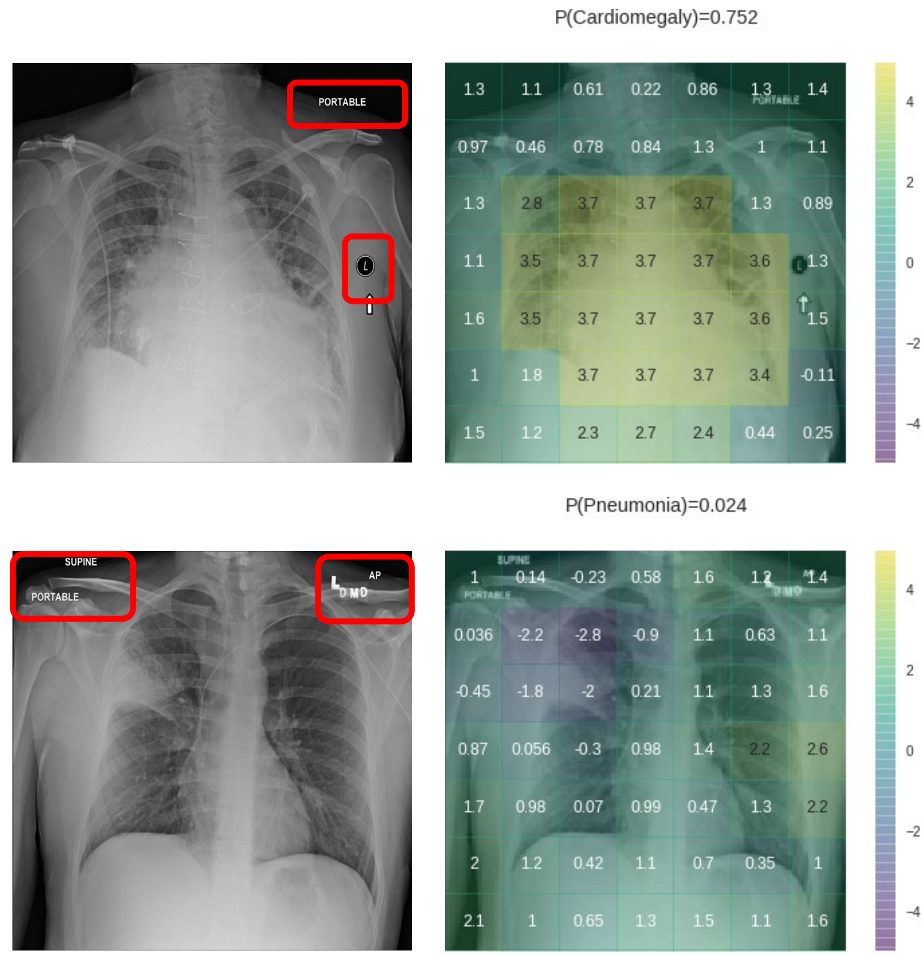
1. Governance options
2. **Disclosure (Transparency)**
3. Auditing
4. Alternatives

Interpretability

- What makes an algorithm *interpretable*?
  - Ability to scrutinize source code?
  - Understanding results from algorithm?
  - Understanding general “logic” of the algorithm?
- Consider initial version of NYC Council bill
  - Required city agencies to publicly release the **source code** of algorithms they use to make their decision
  - In Assignment #1, we start by asking you to consider potential bias of algorithm by just scrutinizing the code
  - How feasible do you think that is?
- And now, it’s *Possibly Apocryphal Story Time*!
  - Show me the tanks!

Totally Verifiable Story Time

- Zech *et al* (2018) examine deep learning models for making disease diagnoses from chest x-rays
    - Neural network is weighting features of x-ray machine
    - Portable machines are more likely to be used with sick patients (who can't come to the hospital)
    - This creates a confound for trying to deploy real systems and can lead to misleading results



Source: John Zech (2018), "What are radiological deep learning models actually learning?", *Medium*, pictures on the left altered by adding red boxes for emphasis. Picture on the right altered by creationg heat maps. "To create the heatmaps, [Zech] use[s] the activations of Zhou et al 2015 and convert them into a probability for each of the 7x7 subregions as described in [his] preprint. I then calculate ln(p_subregion / p_baseline) for each of the 7x7 subregions, where p_subregion is the probability of disease based on that subregion and p_baseline is the population baseline probability of the disease." See this pre-print for reference: https://arxiv.org/pdf/1807.00431.pdf

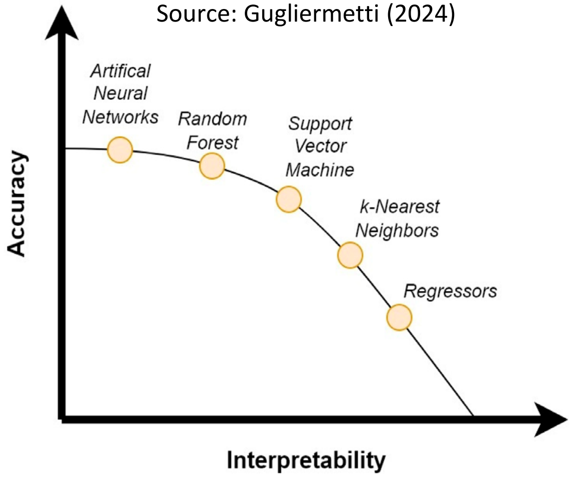
Interpretability vs. Accuracy

- How do we trade-off interpretability vs. accuracy?
  - Generally, more complicated model can be more accurate
  - Some models are more interpretable than others (e.g., a linear function, versus a set of rules, versus a neural network)
  - Hard/impossible for human to understand sufficiently complex models
    - It’s essentially a “opaque box” even if we have access to the code
  - Sometimes, we learn a simpler (or more readily understood) model to approximate the function in a complex model

- How should we weigh interpretability?
  - Importance of transparency and due process
  - E.g., decisions in the legal system

- How should we weigh accuracy?
  - Importance of outcomes
  - E.g., product recommendations

Source: Gugliermetti (2024)



Interpretability via Outcomes

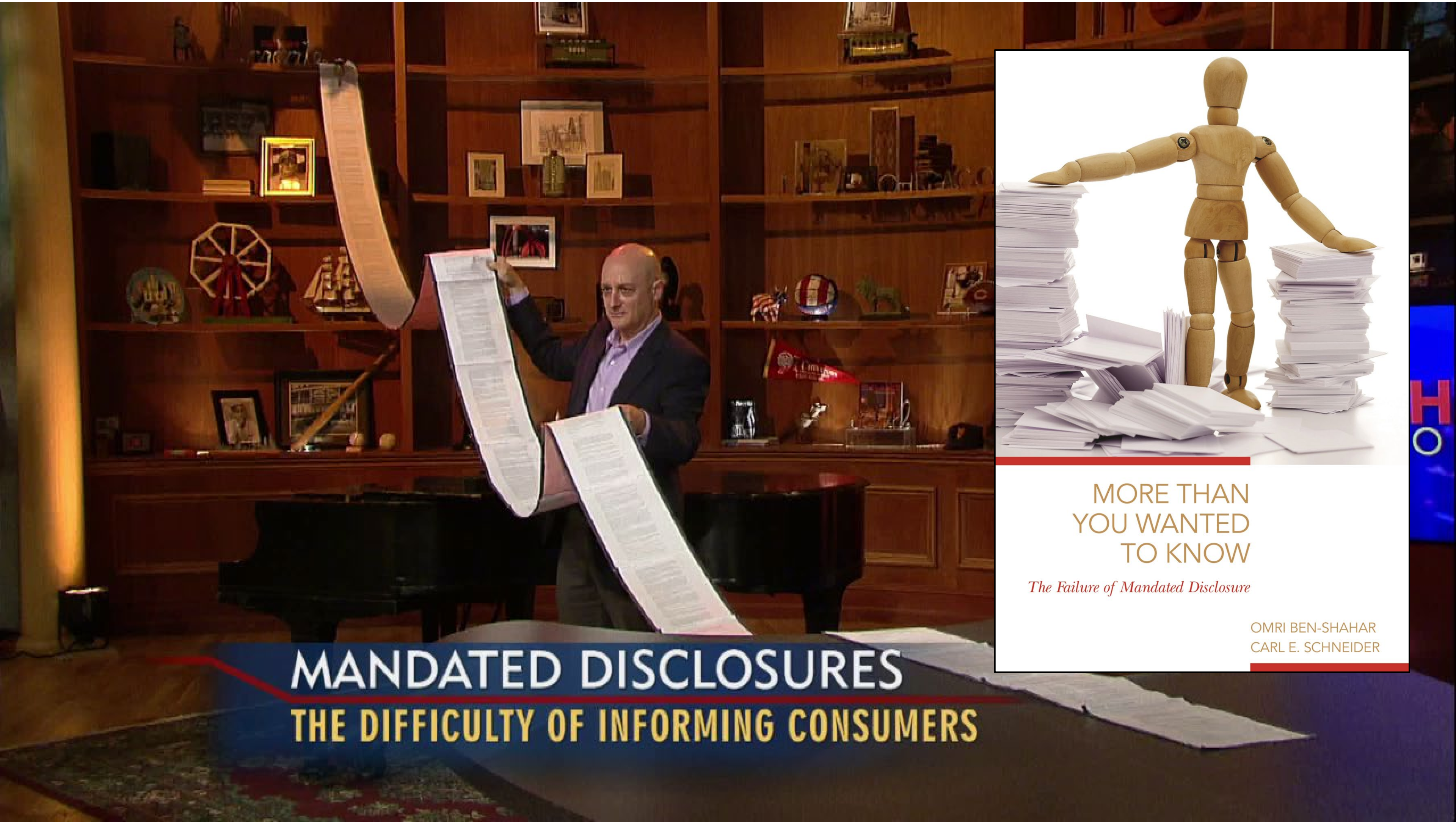
- Accounting for outcomes (outcome-based explanation)
  - How particular inputs lead to particular outputs
- Legal requirement for credit scoring by Fair Credit Reporting Act (FCRA)
  - E.g., You have a right to know factors that impact your FICO score
- Adverse action notice
  - When credit denied, you must be given specific reason for denial
  - Idea is to alert and <u>educate</u> the consumer about credit
- Can’t make credit decision based on protected attributes
  - But, proxy attributes can still be possible inputs (even in explanation)
- Should this apply broadly?
  - *New York Times*, Jan. 21, 2026

Job Applicants Sue to Open ‘Black Box’ of A.I. Hiring Decisions

A recently filed lawsuit claims the ratings assigned by A.I. screening software are similar to those of a credit agency and should be subject to the same laws.

Interpretability via Logic

- Logic of decision-making (logical explanation of result)
  - Need to provide description of "rules" in the system
- European General Data Protection Regulation (GDPR) requires access to "meaningful information about the logic involved" in automated decision-making that impact data subject
  - In case you were wondering, **you** are the "data subject"
  - More about GDPR later
- Seems more transparent, but potentially more problematic
  - What is suitable "logic" for decision?
  - Can someone explain who can see my Facebook post if I post a picture tagging a friend who only allows her friends to see posts on her feed, but I let anyone see mine?
  - How about why Google showed me a particular set of search results?

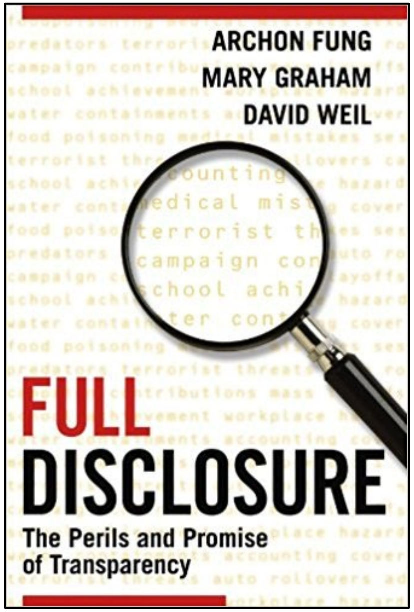

MANDATED DISCLOSURES
THE DIFFICULTY OF INFORMING CONSUMERS
MORE THAN YOU WANTED TO KNOW
The Failure of Mandated Disclosure
OMRI BEN-SHAHAR
CARL E. SCHNEIDER

Source: PBS

iTunes

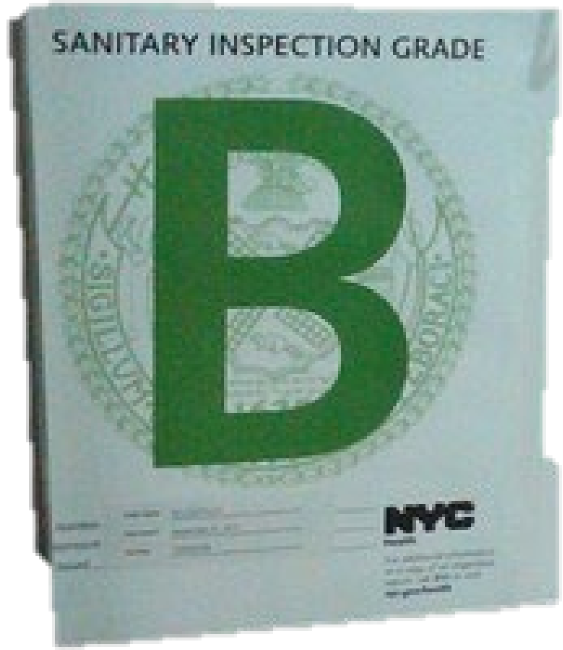
"30-feet long monster of eight-point font"

Source: Omri Ben-Shahar



"highly effective"


COUNTY OF LOS ANGELES
Public Health
ENVIRONMENTAL HEALTH
A
THIS ESTABLISHMENT RECEIVED A SCORE OF
90-100%
AT THE TIME OF INSPECTION
The most recent food inspection report is
available here for review upon request
DO NOT REMOVE UNDER PENALTY OF LAW


SANITARY INSPECTION GRADE
B
NYC


SANITARY INSPECTION GRADE
BEST
NYC

Randomized Controlled Trial

“We find that the grading system had no appreciable effects on foodborne illness, hospitalization, or food handling practices…”

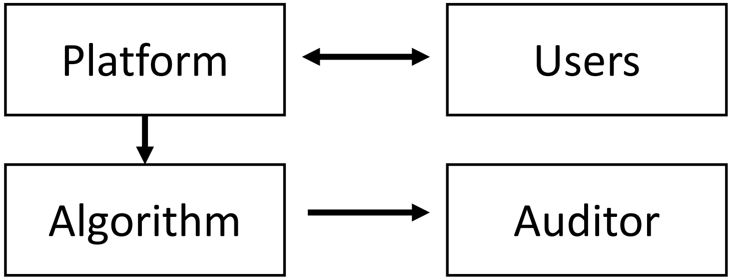
Today's Agenda

1. Governance options
2. Disclosure
3. **Auditing**
4. Alternatives

Auditing Algorithms

- Basic idea of audit: field study to diagnose harmful discrimination/impacts from a decision-making process
  - E.g., Send the same resume with “male” or “female” name to see if there is a difference in interview rates
- Note: algorithms are embedded in software systems/platforms
  - In many cases, we are really auditing the *platform*, not just a single algorithm in it, which can create more complexity
- Auditing of algorithms/platforms can take different forms
  - We’ll consider five from Sandvig *et al* (2014): “Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms”

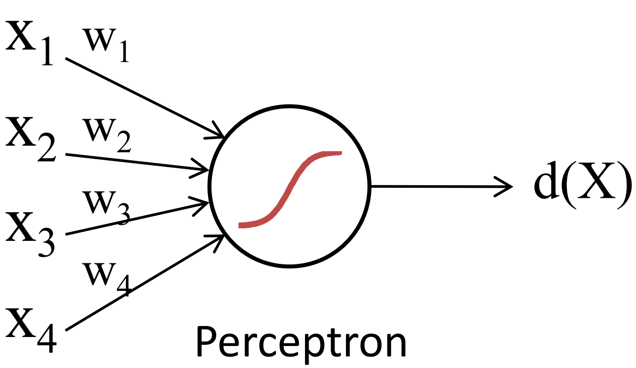
1. Code Audit



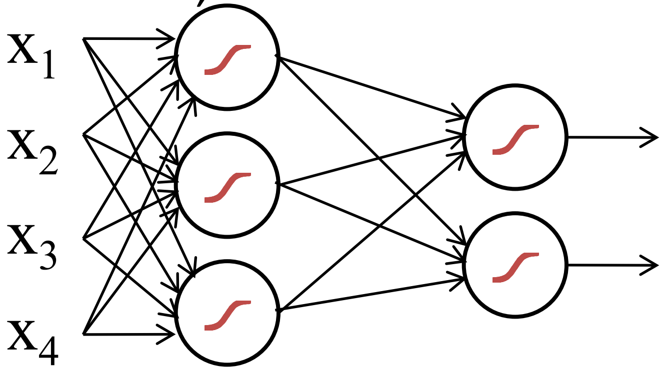
- Platform makes underlying algorithm available to auditor
  - Could open source code and/or model parameters (everyone is auditor)
  - Could put algorithm in safe escrow (designated auditors)
- Provides algorithmic transparency at source code level
  - Essentially, what question 1 of your assignment is about
- Doesn’t necessarily provide data algorithm was trained on
  - That’s why your assignment doesn’t stop after question 1
  - Providing training data can be a privacy concern
    - We’ll talk much more about data privacy in the next unit

Some Machine Learning Architectures

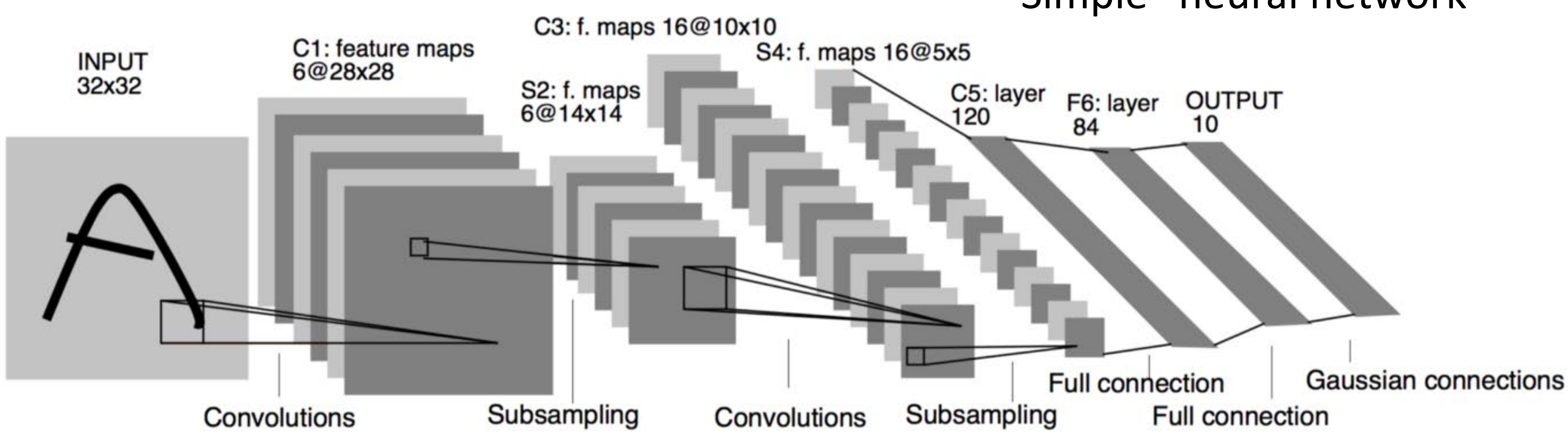
- Accessing algorithm ≠ understanding how decisions are made
  - High-dimensional, non-linear, non-monotonic, discontinuous functions



Perceptron



"Simple" neural network

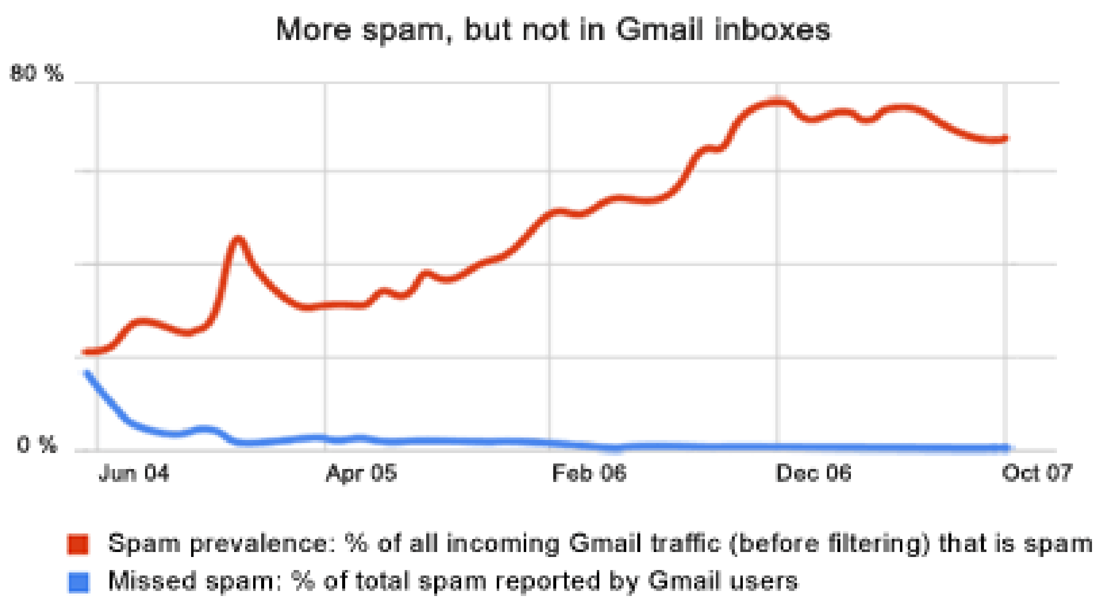


LeNet-5 Deep Neural Network architecture

Source (bottom image): Lecun, Y., et al., Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998: p. 2278--2324.

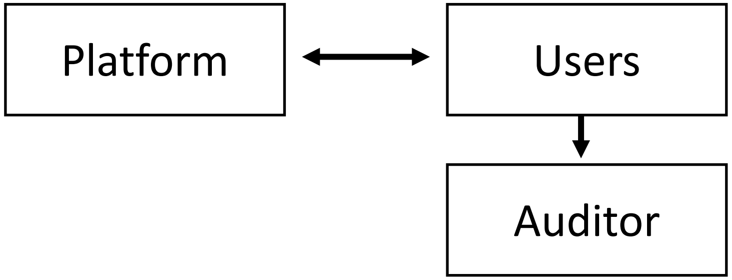
Proprietary Algorithms

- Algorithm can be core intellectual property of company
  - E.g., Email spam filter, Google ranking algorithm
- Release of algorithm could be untenable
  - Gaming/spamming of platform
  - Release of trade secrets to competition
  - Outright theft/copying



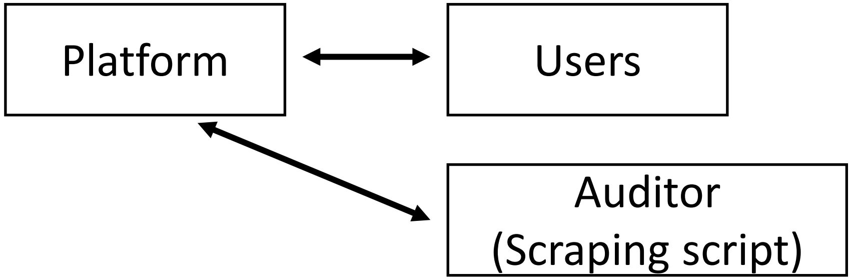
Source: Todd Jackson (October 2007), "How our spam filter works," Official Gmail Blog, https://gmail.googleblog.com/2007/10/how-our-spam-filter-works.html

2. Noninvasive User Audit



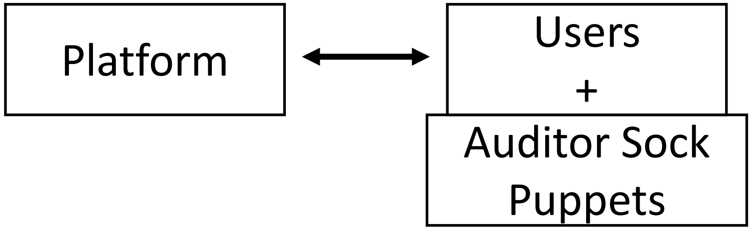
- Auditors ask users to give them results of using platform
- Generally, not randomized in condition assignment
  - Difficult to focus on what you believe may be discriminatory
- Sampling bias
  - Which users you have access to
  - Which users agree to share results with you
  - (Very) limited sample size
- Results from platform may be too sensitive in some domains
  - Finance, healthcare, criminal justice, etc.

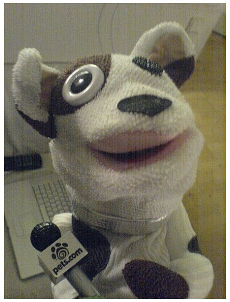
3. Scraping Audit



- Auditors scraps results from platform using program/script
- Potentially violates US Computer Fraud and Abuse Act (CFAA)
  - Criminalizes unauthorized access to computer systems
- Likely to violate platform’s “Terms of Service”
- Should platforms provide an API (or modify their terms of service) to enable this?
  - How do you protect the platform from abuse?
  - Was Cambridge Analytica just auditing Facebook?

4. Sock Puppet Audit





- Auditors create "false users" (sock puppets) to interact with platform
  - E.g., Has been used to test if there is bias in Facebook ad targeting
- Does it violate US Computer Fraud and Abuse Act (CFAA)?
  - Injecting "false" data into system (e.g., fake profiles in LinkedIn)
- Likely to violate platform's "Terms of Service"
- Is this reasonable?
- Will it scale?

Image Source: Wikimedia: https://commons.wikimedia.org/wiki/File:Pets.com_sockpuppet.jpg (CC BY 2.0)

Court Ruling

Creation of Fake Online Accounts to Study Algorithmic Bias Does Not Violate the Computer Fraud and Abuse Act, D.C. Court Rules

By Danielle J. Moss, Joseph O'Keefe and Tony S. Martinez on April 10, 2020

A federal judge recently held that researchers who violate a website's terms of service by creating fake online accounts in order to study algorithmic bias in artificial intelligence software do not violate the Computer Fraud and Abuse Act ("CFAA").

…

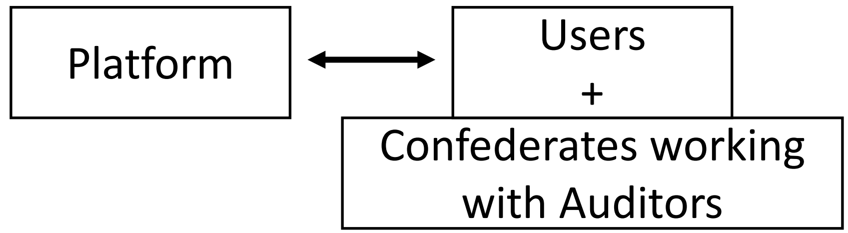
Several of the researchers had created fake employer and employee accounts in order to study the algorithms used by various websites (such as LinkedIn). ***The researchers were particularly interested in determining whether the algorithms discriminated against applicants on the basis of protected characteristics, such as sex, age, or race.***

…

While the Court's decision permits researchers to create fake accounts to study algorithmic bias without concern of criminal liability under the CFAA, it does not absolve individuals of liability arising under other federal and state laws (let alone attorney ethics rules), or shield them from suits by website owners.

Source: Creation of Fake Online Accounts to Study Algorithmic Bias Does Not Violate the Computer Fraud and Abuse Act, D.C. Court Rules byDanielle J. Moss, Joseph O'Keefe and Tony S. Martinez on April 10, 2020 in Proskauer, Law and the Workplace. https://www.lawandtheworkplace.com/2020/04/creation-of-fake-online-accounts-to-study-algorithmic-bias-does-not-violate-the-computer-fraud-and-abuse-act-d-c-court-rules/

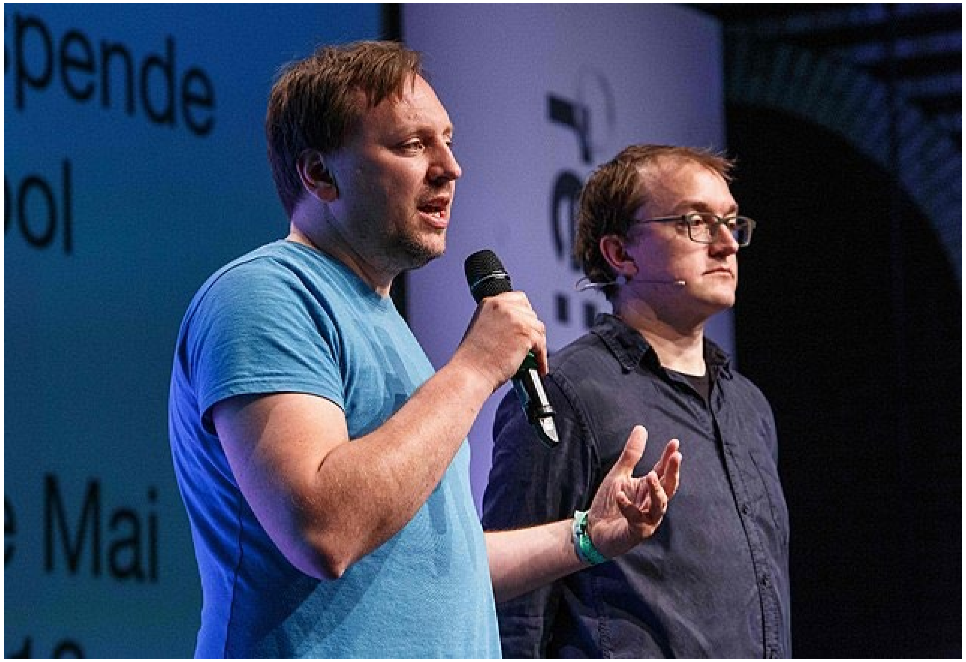
5. Crowdsourced Audit



- Auditors crowdsource real users to interact with platform
- Likely does **not** violate US Computer Fraud and Abuse Act
- Likely does **not** violate platform’s “Terms of Service”
- Does paying confederates change things?
- What if confederates were paid to click on ads?

Crowdsourced Auditing in Action

- TechCrunch, Jan 15, 2019
- Activists in Germany crowdsourcing credit scores (Schufa)
- OpenSchufa: project where users donate their financial data from Schufa
  - 3,000+ data donations

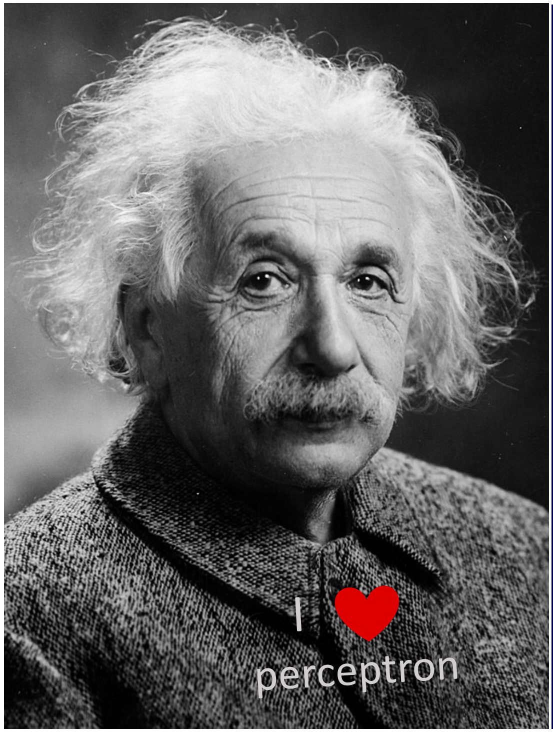


Lorenz Matzat (Idea and Project Lead, OpenSchufa) and Walter Palmetshofer (Concept and realization, OpenSchufa)

Source: re:pubica, Wikimedia Commons (CC-BY 2.0)

Initial results available at:
https://blog.okfn.org/2018/11/29/openschufa-the-first-results/

Changes in the Law



Auditing algorithms. It's not just a good idea…

It's the law!

Source (Albert Einstein): Photograph by Orren Jack Turner, Princeton, N.J, Wikimedia Commons (Public Domain). Modified with Photoshop by PM_Poon and later by Dantadd.

Legally Required Audits

NYC Targets Artificial Intelligence Bias in Hiring Under New Law
By Erin Mulvaney
Bloomberg Law, Dec. 10, 2021

Employers in the city will be banned from using automated employment decision tools to screen job candidates, unless the technology has been subject to a "bias audit" conducted a year before the use of the tool.

…

Companies also will be required to notify employees or candidates if the tool was used to make job decisions.

…

Illinois previously passed a measure similar to New York City's to crack down on the use of such technology in employment decisions.... The attorney general in the Washington, D.C. announced Thursday proposed legislation that would address "algorithmic discrimination" and require companies to submit to annual audits about their technology.

https://news.bloomberglaw.com/daily-labor-report/nyc-targets-artificial-intelligence-bias-in-hiring-under-new-law

Private Initiatives in Auditing

Group Backed by Top Companies Moves to Combat A.I. Bias in Hiring
By Steve Lohr
New York Times, Dec. 8, 2021
…
The Data & Trust Alliance, tapping corporate and outside experts, has devised a 55-question evaluation, which covers 13 topics, and a scoring system. The goal is to detect and combat algorithmic bias.

"This is not just adopting principles, but actually implementing something concrete," said Kenneth Chenault, co-chairman of the group and a former chief executive of American Express, which has agreed to adopt the anti-bias tool kit.

The companies are responding to concerns, backed by an ample body of research, that A.I. programs can inadvertently produce biased results.

https://www.nytimes.com/2021/12/08/technology/data-trust-alliance-ai-hiring-bias.html

01-26-21 | POV FAST COMPANY

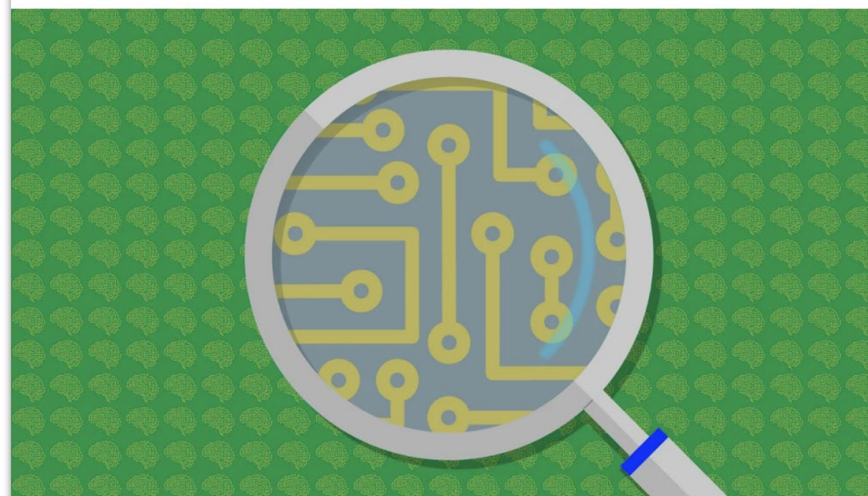
Independent auditors are struggling to hold AI companies accountable

Controversial AI company HireVue implied that an external audit showed its algorithms had no bias. But a look at the audit itself tells a different story.

“[T]here are countless ***subjective choices*** in each specific audit. These choices can easily tip an audit towards a favorable view of the client.”

A new wave of AI auditing startups wants to prove responsibility can be profitable

Businesses hate regulatory compliance, but love profits. AI auditing startups could help them accomplish both.



"[T]he status quo system was largely corrupted, with auditors systematically reporting plant emissions just below the standard . . .
Second, [common pool payment] caused auditors to report more truthfully and very significantly lowered the fraction of plants that were falsely reported as compliant . . ."

Source: Duflo et al. (2013)

Today's Agenda

1. Governance options
2. Disclosure
3. Auditing
4. **Alternatives**

Fairness Optimized?

Equal Employment Opportunity Commission Guidance

Selection rate = # hired / # applicants

Selection ratio (between groups)

$$\frac{\text{Selection rate for protected group}}{\text{Selection rate for reference group}} = \frac{\text{1 / 34 for non-white applicants}}{\text{10 / 43 for white applicants}}$$

$$\frac{3\%}{23\%} = 13\% \quad \mathbf{< 80\%}$$

Fairlearn

einstein

Aequitas

Bias & Fairness Audit

pymetrics/**audit-ai**

detect demographic differences in the output of machine learning models or other assessments

AI Fairness 360

fAIrplay

U.S. Patent Oct. 8, 2019 Sheet 4 of 7 US 10,438,135 B2

Performance Model Adverse Impact Correction

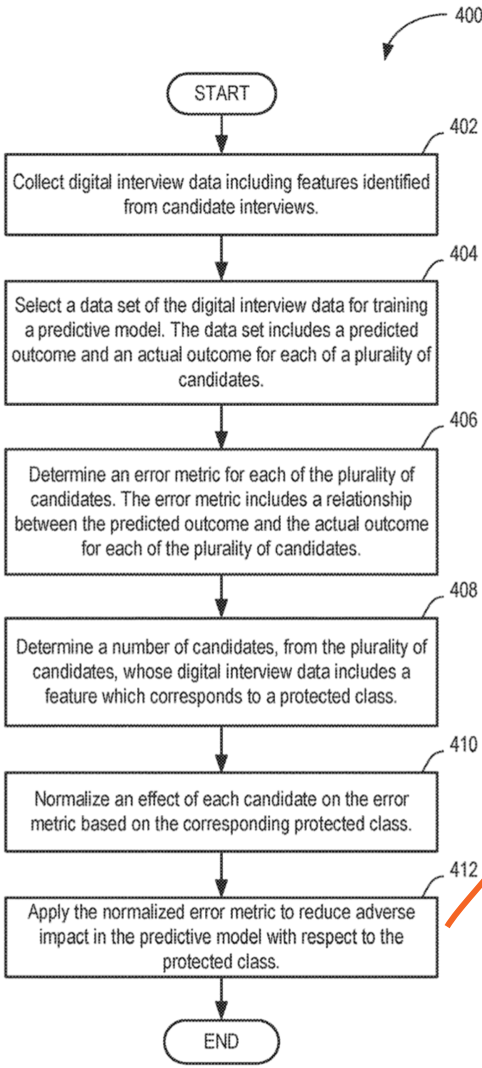
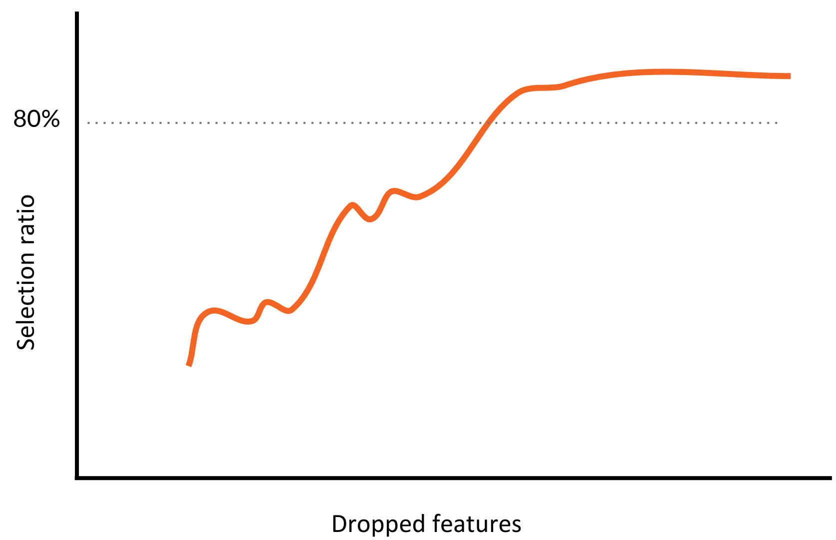


FIG. 4

At block 412, the processing may apply the normalized error metric [e.g., the 80% rule] to reduce adverse impacts . . . with respect to the protected class by directing the model or dropping features. . . .



Data Considerations

Buolamwini & Gebru (2018)

Pilot Parliaments Benchmark

Error rates for darker-skinned females up to 34.7%

Balakrishnan et al. (2021)

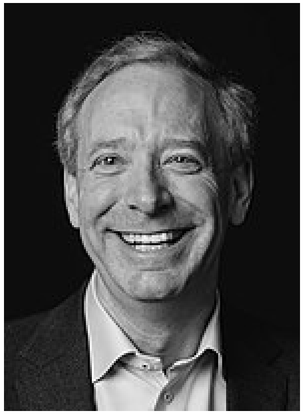
Synthetic Images

“Very few males have long hair and almost no light-skinned females have short hair”

Varying Hair length

Varying Skin Tone

Proposal From Microsoft President



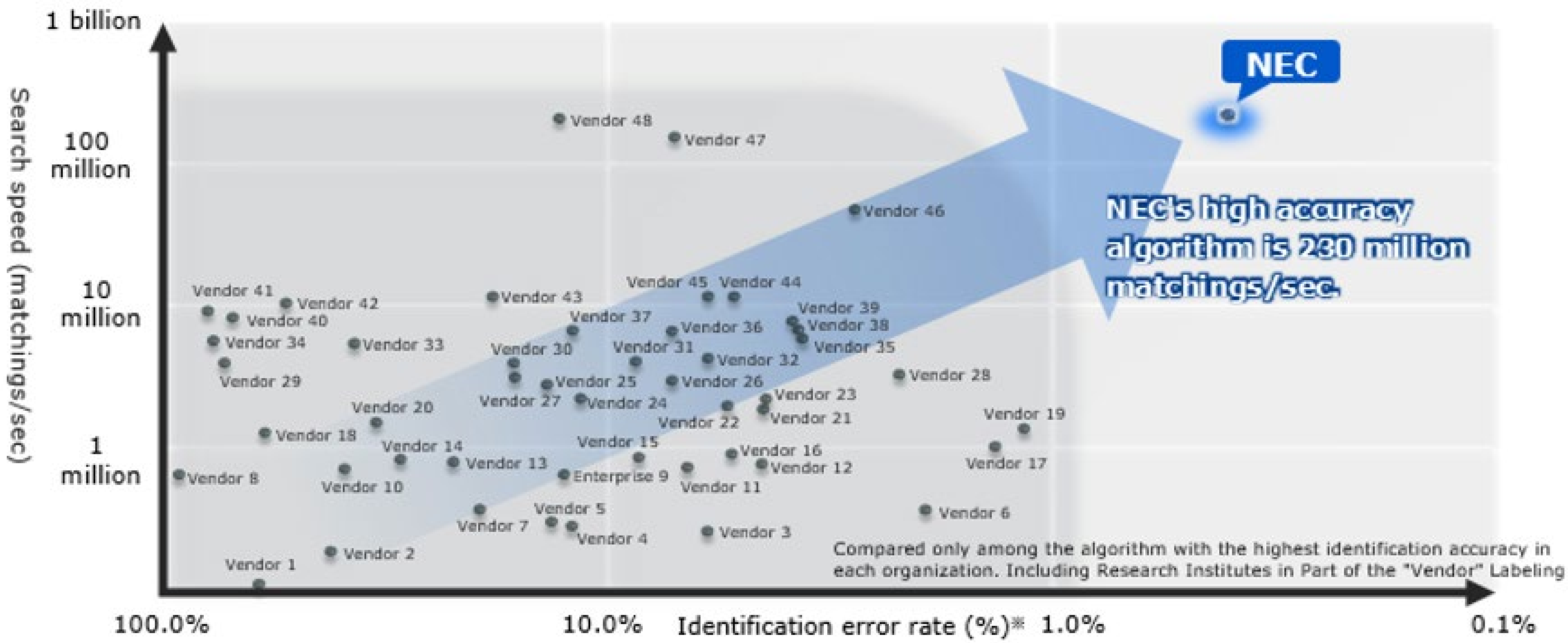
- Brad Smith, President and Chief Legal Officer at Microsoft, has proposed a marketplace solution
- He claims that no one wants unfair algorithms
- Have companies evaluate their algorithms on a standard benchmark using a public dataset and report the results
- Consumers can choose to purchase/use the algorithm they deem best, given fairness results, price, etc.
- Do you think this is a reasonable proposal?

Microsoft improves facial recognition technology to perform well across all skin tones, genders

June 26, 2018 | John Roach

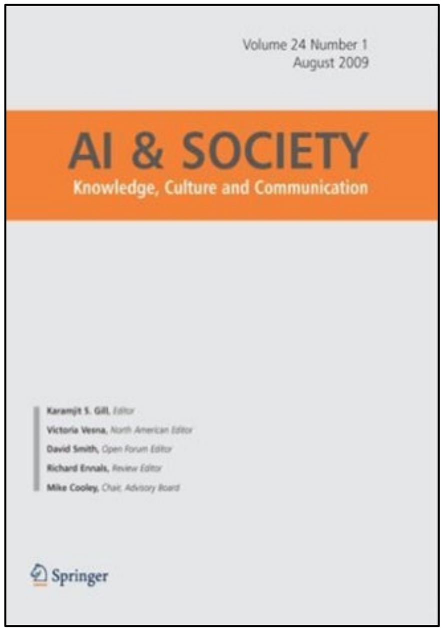
NIST

FRVT
FACE
RECOGNITION
VENDOR TEST



※ False negative identification rate at a false positive identification rate of 0.1% at 1.6 million registered people

Participatory Design



Against "Democratizing AI"

Johannes Himmelreich[1]

"morally myopic"

"AI should be democratized not by broadening and deepening participation but by increasing the democratic quality of the administrative and executive elements of collective decision making."