CMSC 304 Computer Ethics

Dr. Cynthia Matuszek

1

Bookkeeping

- Please continue to journal consistently
  - Lowest 3 journal scores can be dropped
- Questions in EA 2
  - We're using the worksheet to write a short essay
  - Questions in the case study are not part of the assignment
- A note on Discord: primarily for contacting TA/each other

2

Bias in what, exactly?

- AI ≠ Machine learning
  - (And ML ≠ deep learning)
- Much modern AI **depends on** ML
- But exceptions exist
- Bias appears in **models**, which are present in all of AI

3

Algorithmic Bias

- Models can systematically mistreat certain socio-economic groups
- Biased models lead to biased policy decisions
- Need to avoid policies that lead to discrimination
- Example: estimating risk based on race

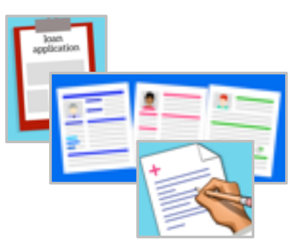
Bias has a technical meaning in ML. That's not what we're talking about.

Slide: knowledge4policy.ec.europa.eu/sites/default/files/2021_EUCONFMOD_Vega.pptx

4

Bias in AI: What is it?

- What is bias in AI?
  - What does it mean?
  - How does it manifest?
- Where does bias come from?
  - Data and models
  - Design decisions
  - Ecosystems and people



pixabay.com/illustrations/loan-agreement-signature-business-4273819/
pixabay.com/illustrations/career-resume-hiring-job-interview-3449422/
pixabay.com/illustrations/prescription-medical-doctor-pad-4545598/

5

Bias in AI: Considerations

- What are technical solutions?
- What are non-technical considerations?
- Who has responsibilit(ies)?
- How serious a problem is it?
- What harm(s) does it do?
  - Ethically
  - Legally

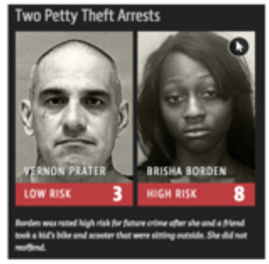
6

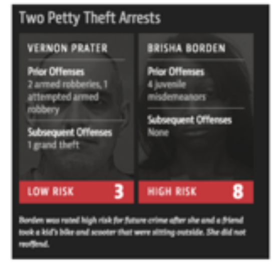
Examples

- A Google search for Jew returns anti-Semitic web pages
- Google's image tagging algorithm tags black people as 'Gorilla'
- Men working in residential kitchens in images are labeled women
- A Nikon Coolpix S630 asks if someone blinked on Asian faces
- An image search for "beautiful" returns all white people
- Machine-translated Turkish inserts gendered roles for pronouns
- 28 members of Congress incorrectly recognized by facial recognition software trained on a database of mug shots
- Amazon's resume evaluation software excludes resumes from women
- …and on and on

7

Some examples: Sentencing





https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

11

Legally Recognized Protected Classes

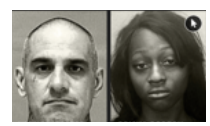
- **Race** (Civil Rights Act of 1964); **Color** (Civil Rights Act of 1964); **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964); **Religion** (Civil Rights Act of 1964);**National origin** (Civil Rights Act of 1964); **Citizenship** (Immigration Reform and Control Act); **Age** (Age Discrimination in Employment Act of 1967); **Pregnancy** (Pregnancy Discrimination Act); **Familial status** (Civil Rights Act of 1968); **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990); **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

Sarah Bird, Developing AI (Responsibly

20

Types of Harm

- Harms of allocation: withhold opportunity or resources
- Harms of representation: reinforce subordination along the lines of identity, stereotypes



Sarah Bird, Developing AI (Responsibly
Cramer et al 2019, Shapiro et al., 2017, Kate Crawford, "The Trouble With Bias" keynote NeurIPS'17

21

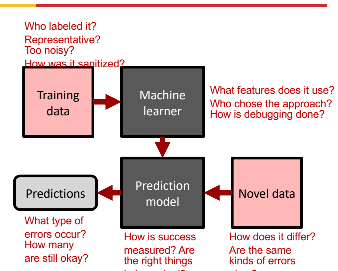
Accuracy vs. Fairness?

- Probably not!
  - Many definitions of fairness
  - Prejudicial results are a form of inaccuracy
- Making a system "fair" does not mean making it less good

22

It's Not Just the Data

- ML bias is not just a data problem
  - Design assumptions
  - Testing metrics
  - Makeup of the development team
  - Decisions about error type tradeoffs
  - Reporting of failures (or lack thereof)
  - And, yes, data representativeness



23

Negative Legacy

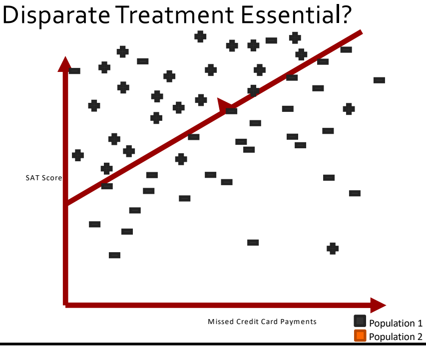
- When algorithms accurately capture the bias present in existing sources
- Codifies decades of systemic bias
- Difficult to remove
- Difficult to identify all relevant factors
- Redlining; “executed”
- Removing information about class membership can make it **worse**

https://www.infoworld.com/article/3607748/3-kinds-of-bias-in-ai-models-and-how-we-can-address-them.html

24

Is Disparate Treatment Essential?

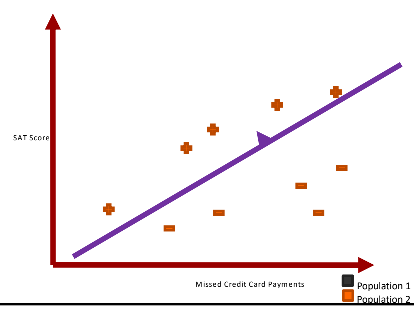
Aaron Roth, (un)fairness in Machine Learning



25

Is Disparate Treatment Essential?

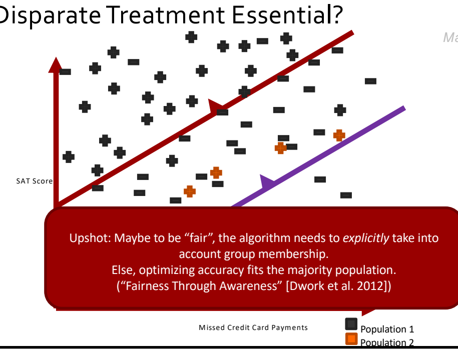
Aaron Roth, (un)fairness in Machine Learning



26

Is Disparate Treatment Essential?

Aaron Roth, (un)fairness in Machine Learning



27

Underestimation

- Without examples, ML systems don’t provide reliable predictions
- If your workforce is predominantly male, you can’t make predictions about female resumes from the data you have
- One source of bias: you don’t know how well the people would have done that you **didn’t** hire

https://www.infoworld.com/article/3607748/3-kinds-of-bias-in-ai-models-and-how-we-can-address-them.html

28

Sources of Error

- **Sample bias:** a dataset does not reflect the reality
  - Example: facial recognition trained primarily on white men
- **Exclusion bias:** deleting valuable data thought to be unimportant
  - Example: removing geographical data because 98% of customers are in U.S.
- **Recall bias:** label similar types of data inconsistently
  - Example: loan applications labeled by different people as higher or lower risk
- **Observer bias:** Also known as confirmation bias, observer bias is the effect of seeing what you expect to see or want to see in data
  - Example: rejecting female resumes as less qualified

www.telusinternational.com/articles/7-types-of-data-bias-in-machine-learning

29

Types of Sample Bias

- Reporting bias: the frequency of events in the training dataset doesn't accurately reflect reality
  - Example: all customers living in a remote region have a high risk of fraud
  - Explanation: investigators hated traveling there, went only in egregious cases
- Selection bias: training data is either unrepresentative or is selected without proper randomization
  - Example: face recognition performs poorly on Black and female faces
  - Explanation: highly white male-dominated training data

itrexgroup.com/blog/ai-bias-definition-types-examples-debiasing-strategies

30

What is Algorithmic Fairness?

- **Demographic parity:** similar outcomes observed across all groups
  - Ex: 20% of female applicants and 20% of male applicants receive loans
  - The problem: what if groups actually differ? Men default more than women!
- **Equality of false negatives:** enforce consistent rates of false negatives across all groups
  - The percentage of male and female loan applicants that is incorrectly rejected is consistent
  - The problem: evaluating false negatives is sometimes quite difficult
- **Equalized odds:** false negative and true negative rates are consistent across groups

mostly.ai/blog/we-want-fair-ai-algorithms-but-how-to-define-fairness

31

Discussion: computerizing gaydar

- Read http://tiny.cc/gaydar (and do note, you *cannot* tell whether someone is gay from "face structure"—that part is incorrect)

1. What are some potential **harms** and **benefits** from this work?
2. Who are some major stakeholders in the discussion?
   - Authors of the work; Stanford; the APA; who else?
3. Should this work have been performed?
   - If so, why? If not, why not?
4. What policies should exist, at what level, that apply to…
   - Whether the work is performed?
   - What to do with the results?

www.nbcnews.com/feature/nbc-out/controversial-ai-gaydar-study-spawns-backlash-ethical-debate-n801026

32