Lecture 9:
Fairness in AI/ML

CMSC 25900 / DATA 25900

Spring 2021

The University of Chicago

Machine Bias (ProPublica)

- COMPAS System for risk assessment
- Based on answers to 137 questions
- ProPublica obtained data:
  - Broward County, Florida
- "And it's biased against blacks."
  - Northpointe: It's equally accurate across demographic groups!

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.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.
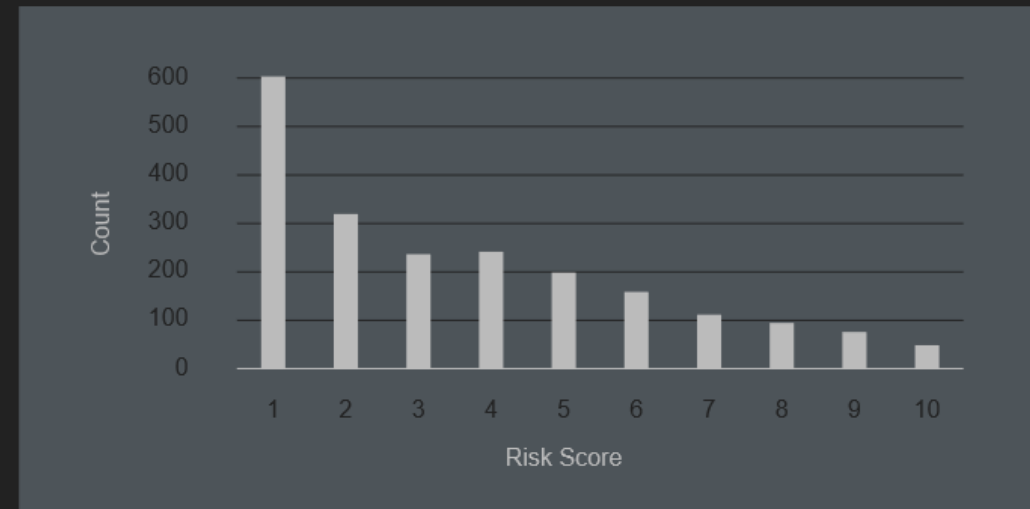
COMPAS

- Evidence of discrimination?

Black Defendants' Risk Scores
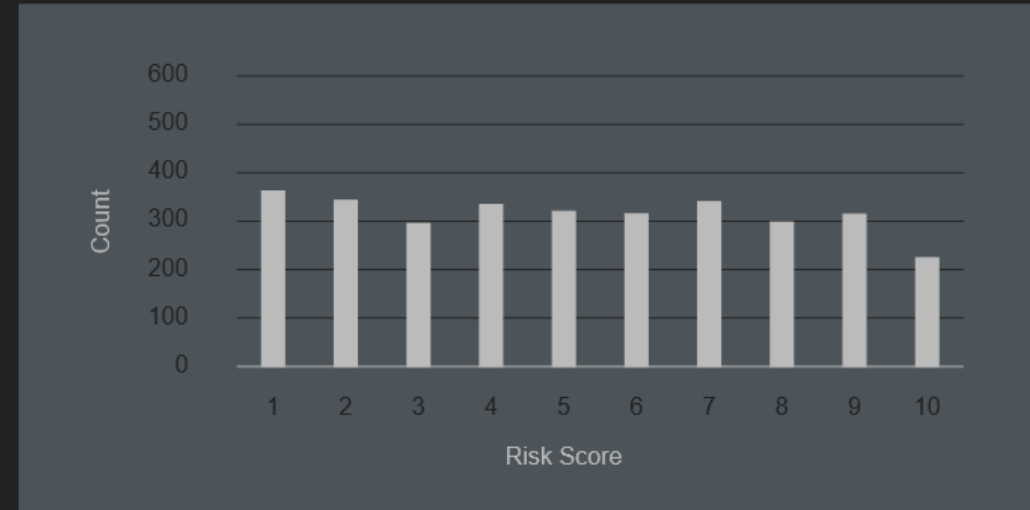
White Defendants' Risk Scores

These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. (Source: ProPublica analysis of data from Broward County, Fla.)

COMPAS

- Evidence of discrimination?

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

Proposal: Algorithmic Grading in 259

- The data we have:

Name	Age	Department	Gender	Grade in CS 119	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 1

- Let's extrapolate from the Assignment 1 grade

Name	Age	Department	Gender	Grade in CS 119	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 1

- Small data! We also advertised something different!

Name	Age	Department	Gender	Grade in CS 119	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 2

- Let's extrapolate from the CS 119 grade

Name	Age	Department	Gender	Grade in CS 119	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 2

- Is this just? Does Jane get a grade?

Name	Age	Department	Gender	Grade in CS 119	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 3

- Let's use Department and the Grade in CS 119

Name	Age	Department	Gender	Grade in CS 119	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 3

- Why should these matter?

Name	Age	Department	Gender	Grade in CS 119	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 4

- Let's use all demographics and the Grade in CS 119

Name	Age	Department	Gender	Grade in CS 119	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 4

- Why?!?! (Also, age and gender are protected classes)

Name	Age	Department	Gender	Grade in CS 119	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 5

- Everyone gets an A!

Name	Age	Department	Gender	Grade in CS 119	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 6

- Everyone gets an F!

Name	Age	Department	Gender	Grade in CS 119	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 6

- Societal notions of justice may imply that failing everyone is bad

Name	Age	Department	Gender	Grade in CS 119	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

The Difficulty of Fairness

- Terminology is conflated across disciplines
  - Political philosophy
  - Employment lawyer
  - Computer scientist

- See: Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, Richmond Y. Wong. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. PACM HCI (CSCW), 2019.

Individual Fairness

- One of the early definitions of fairness
- **Individual fairness:** Similar people should be treated equally

Statistical Non-Discrimination

- Basis in employment and housing law (e.g., Fair Housing Act)
- Primarily considers *protected classes*
  - Race, gender, sex, national origin, religion, marital status, etc.
- In this approach to fairness, we want to approximately equalize some quantities across demographic groups (**group fairness**)
  - Mainly focuses on **disparate impact** (treating different groups differently)

Group Fairness (Just a Few Approaches)

- Demographic parity (equal outcomes)
  - Equalize the chance of positive classifications across groups

Group Fairness (Just a Few Approaches)

- Equalizing accuracy across groups

Group Fairness (Just a Few Approaches)

- Equalized odds (true positive rate and false positive rate are equal)

Blindness to Protected Classes

- Should we just intentionally not collect data about whether or not data subjects belong to a protected class?
    - The answer is very complicated. It's often (but not always!) "no"… why not?

How Does Sampling Impact Fairness?

How Does Sampling Impact Fairness?

- What if our sample is unbalanced? Can that cause problems?
- What if our sample is not representative?
- What if we collect the wrong features?

Concept Drift – The Passage of Time

- Can we be embedding historical biases?

Time
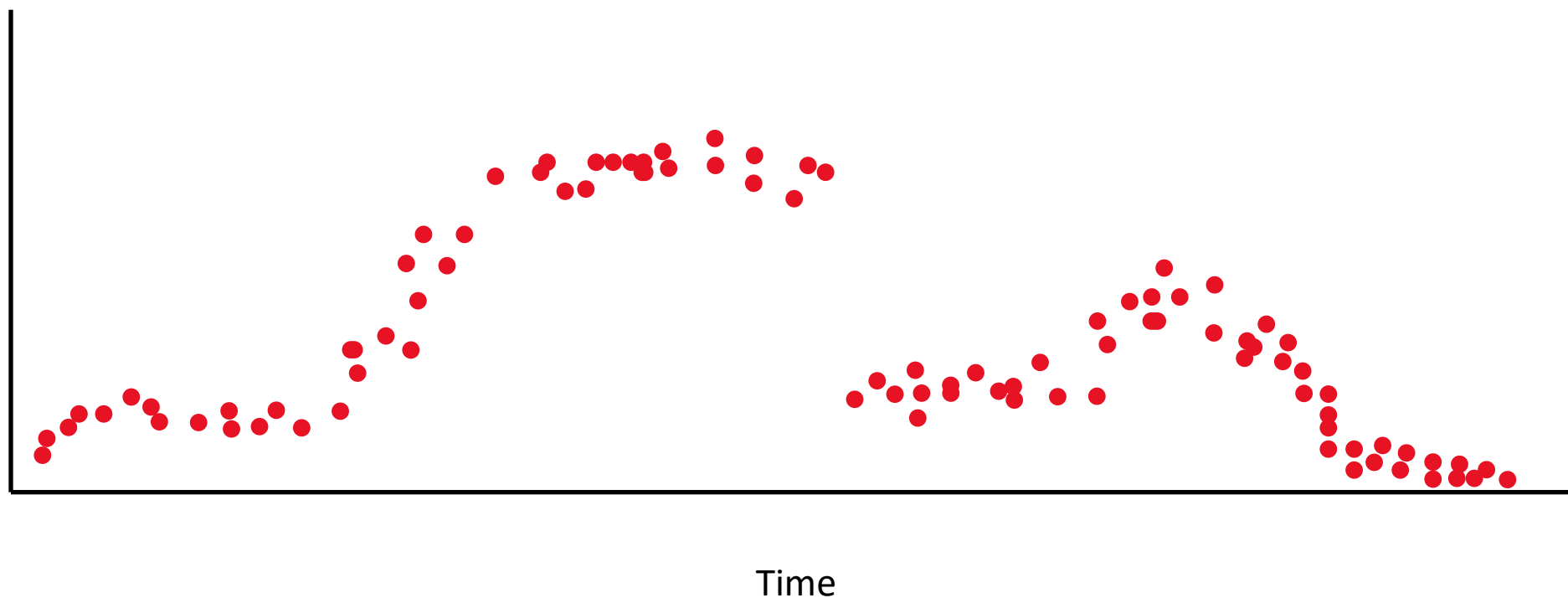
Process Fairness

- How do we decide what predictor variables to include?

- **Process fairness**: Exclude from the model predictor variables that are _deemed to be_ unfair for the classification task

- Should we just crowdsource perceptions?
    - Grgic-Hlaca et al. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In _Proc. WWW_, 2018.
    - Important question: Who gets to decide what is fair? Is it majoritarian voting? Should it be experts in law/technology?

Some Attempted Fairness Mitigations

- Transform the training data features and/or labels
- Change the weights in the model produced
- Adversarial de-biasing
  - e.g., using a discriminator from a Generative Adversarial Network

AI Fairness 360

- IBM open source project: https://aif360.mybluemix.net/
- Online demo: https://aif360.mybluemix.net/data
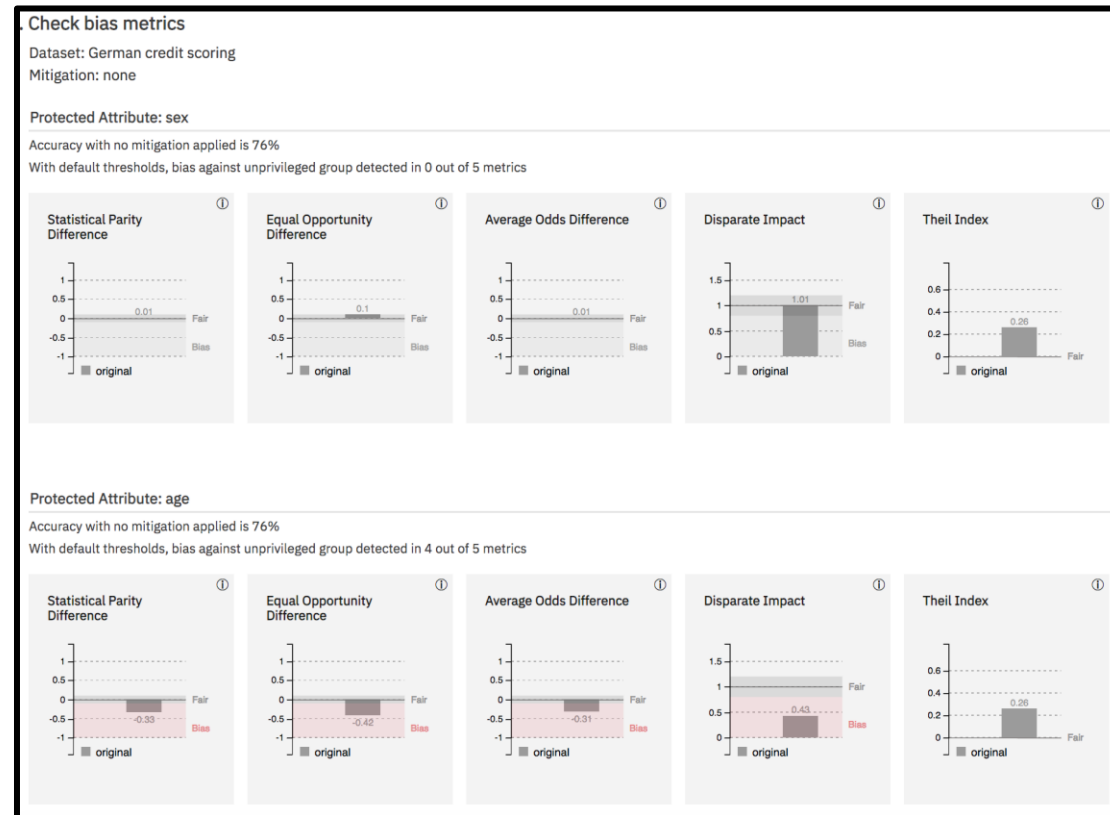
What-If Tool

- Google open source project: https://pair-code.github.io/what-if-tool/
- Online demo: https://pair-code.github.io/what-if-tool/image.html

What-If Tool

Aequitas Tool

- UChicago open source project: http://www.datasciencepublicpolicy.org/projects/aequitas/
- Online demo: http://aequitas.dssg.io/example.html

Audit Results: Bias Metrics Values

race

Attribute Value	False Discovery Rate Disparity	False Positive Rate Disparity	False Negative Rate Disparity
African-American	0.91	1.91	0.59
Asian	0.61	0.37	0.7
Caucasian	1.0	1.0	1.0
Hispanic	1.12	0.92	1.17
Native American	0.61	1.6	0.21
Other	1.12	0.63	1.42

Counterfactuals and Recourse

- **Counterfactual:** Ideally small difference(s) in a data subject's set of features that would cause a different classification
  - Need a distance metric! But not all variables are created equal.
- **Recourse:** The ability for a data subject to change particular predictor variables
  - Contrast using "the timeliness of credit card payments" versus "the number of years of credit history" versus "sex"
  - To what extent should models **nudge** (influence, but not force) particular behavior?

Unsupervised Models Are Biased, Too!

- https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html?m=1

As Machine Learning practitioners, when faced with a task, we usually select or train a model primarily based on how well it performs on that task. For example, say we're building a system to classify whether a movie review is positive or negative. We take 5 different models and see how well each performs this task:
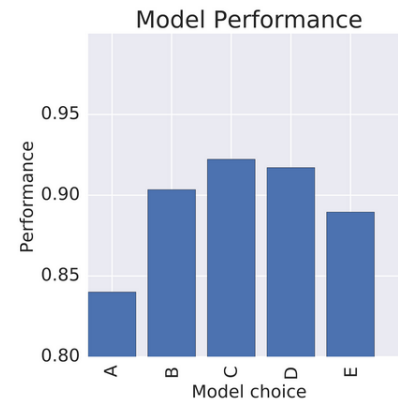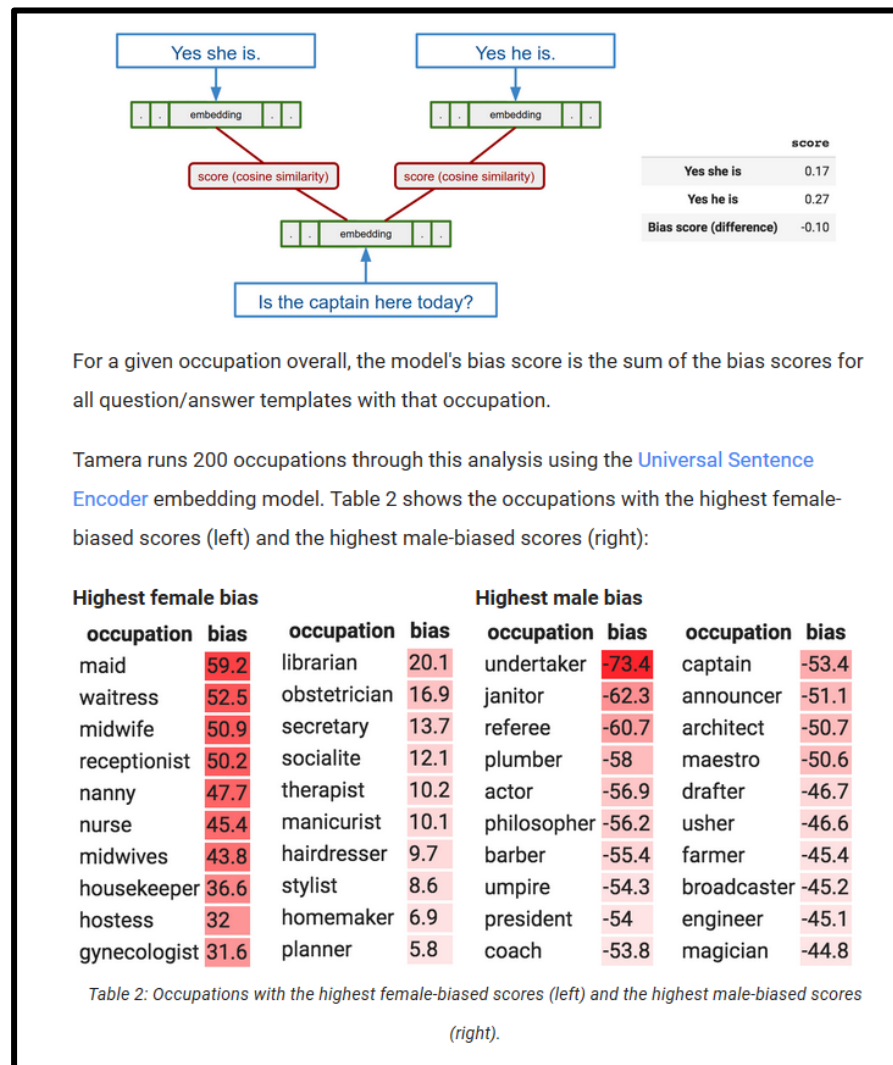
Model Performance

Figure 1: Model performances on a task. Which model would you choose?

Normally, we'd simply choose Model C. But what if we found that while Model C performs the best overall, it's also most likely to assign a more positive sentiment to the sentence "The main character is a man" than to the sentence "The main character is a woman"? Would we reconsider?

Gender Biases of Chatbots

Table 2: Occupations with the highest female-biased scores (left) and the highest male-biased scores (right).

Word Embeddings

Text embedding models convert any input text into an output vector of numbers, and in the process map semantically similar words near each other in the embedding space:

| 0.4 | -0.2 | ... | ... | ... | ... | ... | ... | 1.7 |

that rug really tied the room together

- go to work
- commute
- horse
- dog
- cat
- we're in the garden
- he used the shovel
- the building is tall
- please use the rake
- the house is big
- I used the hose

Gender Biases of Chatbots

Targets (N)	Attributes (N)	GloVe*	word2vec	nnlm-en-dim50	nnlm-en-dim128	universal
Flowers vs Insects (25)	Pleasant vs Unpleasant (25)	1.50*	1.54*	1.54*	1.63*	1.38*
Instruments vs Weapons (25)	Pleasant vs Unpleasant (25)	1.53*	1.63*	1.66*	1.55*	1.44*
Eur-American vs Afr-American Names[6] (25)	Pleasant vs Unpleasant[6] (25)	1.41*	0.58*	0.70*	0.04	0.36
Eur-American vs Afr-American Names[7] (18)	Pleasant vs Unpleasant[6] (25)	1.50*	1.24*	1.04*	0.23	-0.37
Eur-American vs Afr-American Names[7] (18)	Pleasant vs Unpleasant[8] (8)	1.28*	0.72*	0.28	-0.09	0.72
Male vs Female names (8)	Career vs Family (8)	1.81*	1.89*	1.45*	1.70*	0.03
Math vs Arts (8)	Male vs Female (8)	1.06	0.97	1.29*	1.07	0.59
Mental vs Physical Disease (6)	Temporary vs Permanent (7)	1.38*	1.30	1.35*	0.96	1.60*
Science Arts (8)	Male vs Female (8)	1.24*	1.24*	1.34*	1.19	0.24
Young vs Old Names (8)	Pleasant vs Unpleasant (8)	1.21	-0.08	0.75	-0.47	1.01

Table 1: Word Embedding Association Test (WEAT) scores for different embedding models. Cell color indicates whether the direction of the measured bias is in line with (blue) or against (yellow) the common human biases recorded by the Implicit Association Tests. *Statistically significant (p < 0.01) using Caliskan et al. (2015) permutation test. Rows 3-5 are variations whose word lists come from [6], [7], and [8]. See Caliskan et al. for all word lists. * For GloVe, we follow Caliskan et al. and drop uncommon words from the word lists. All other analyses use the full word lists.

https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html?m=1

Reconceptualizing Fairness as Justice

- Should we follow Rawls and consider justice as fairness?
- Should we start thinking about fairness in terms of trolley problems? https://en.wikipedia.org/wiki/Trolley_problem
- How might our societal notions of what is just change how we build a classifier, **as well as whether we use ML at all**?
- How do we think about due process within fairness?
- Returning to the COMPAS example: How did judges use (or choose not to use) COMPAS risk scores? Is this just?
- Accountability? Transparency? Explanations?

Agree or Disagree?

"An algorithm can't be biased."