

# Bias and Fairness in Machine Learning

---

Irene Y. Chen

 @irenetrampoline



## Beauty contest judged by AI and the robots discriminate against dark skin

3 days ago | Published by : Avinash Nandakumar

HIDDEN BIAS

View more Alarming

# 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.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of \$80.

## Is an algorithm any less racist than a human?

Employers trusting in the impartiality of machines sounds like a good plan to eliminate bias, but data can be just as prejudiced as we are



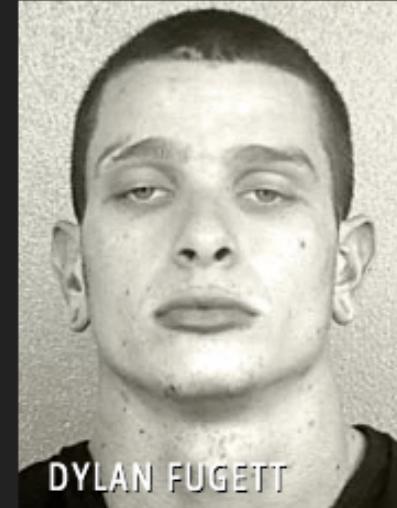
A wealth of startups have sprung up in recent years to address the appetite for more diverse workforces by utilising algorithms. Photograph: Alamy Stock Photo

We would all like to fancy ourselves as eminently capable of impartiality, able to make decisions without prejudices - especially at work. Unfortunately, the reality is that human bias, both conscious and unconscious, can't help but come into play when it comes to who gets jobs and how much money candidates get offered.

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



## Two Drug Possession Arrests



DYLAN FUGETT

LOW RISK

3



BERNARD PARKER

HIGH RISK

10

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

<http://gendershades.org/overview.html>

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

# COMPAS

- ▶ Correctional Offender Management Profiling for Alternative Sanctions
- ▶ Used in prisons across country: AZ, CO, DL, KY, LA, OK, VA, WA, WI
- ▶ “Evaluation of a defendant’s rehabilitation needs”
- ▶ Recidivism = likelihood of criminal to reoffend

# COMPAS (continued)

- ▶ “Our analysis of Northpointe’s tool, called COMPAS (which stands for Correctional Offender Management Profiling for Alternative Sanctions), found that black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk.”

1. **COMPAS analysis**
2. What is fairness in machine learning?
3. Quantitative definitions of fairness in supervised learning
4. Practical tools for analyzing bias
5. Solutions, ethics, and other curveballs

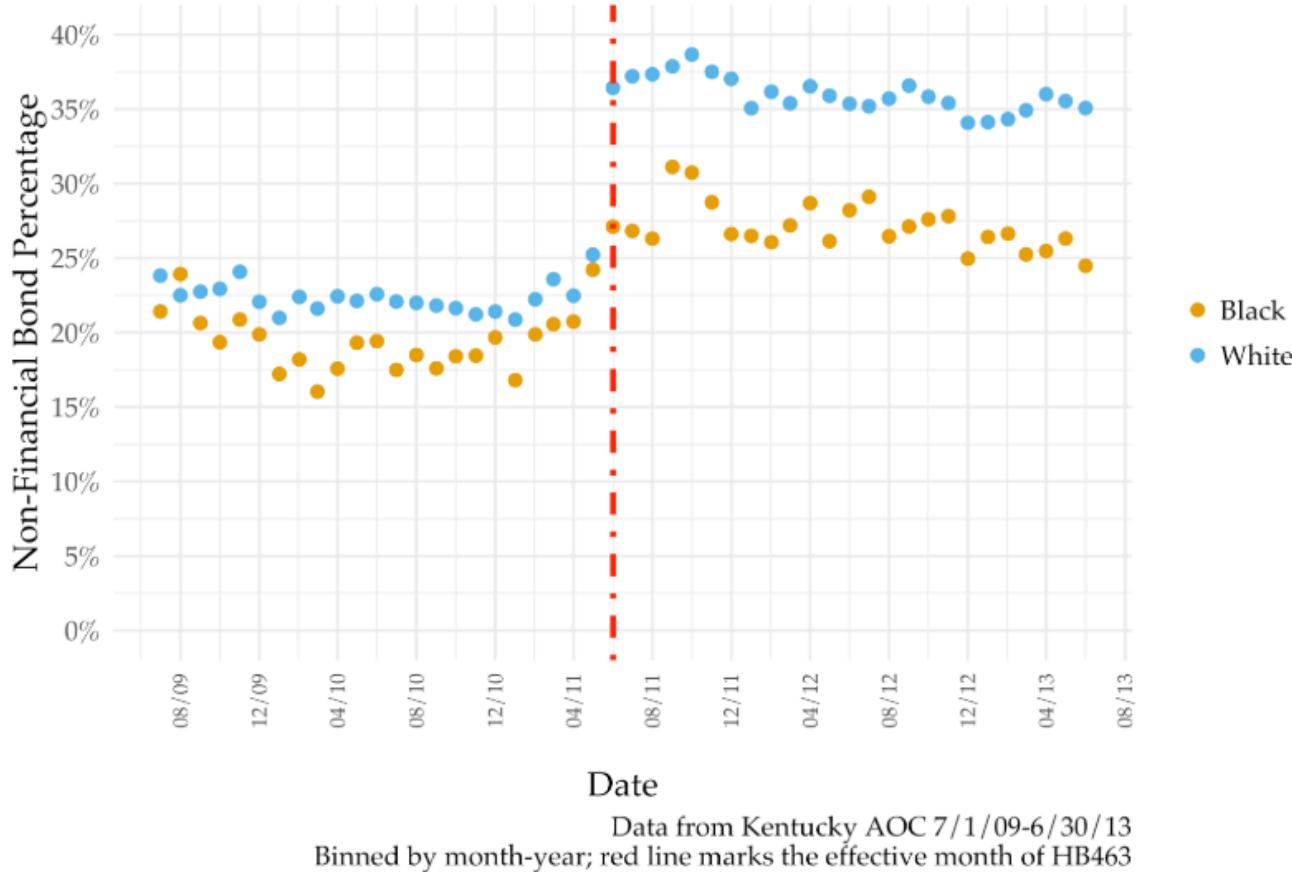
- ▶ Original: <https://github.com/propublica/compas-analysis/blob/master/Compas%20Analysis.ipynb>
- ▶ Exercise: <https://github.com/irenetrampoline/compas-python>
- ▶ Colab solutions: <http://bit.ly/sidn-compas-sol>

# Practicum options

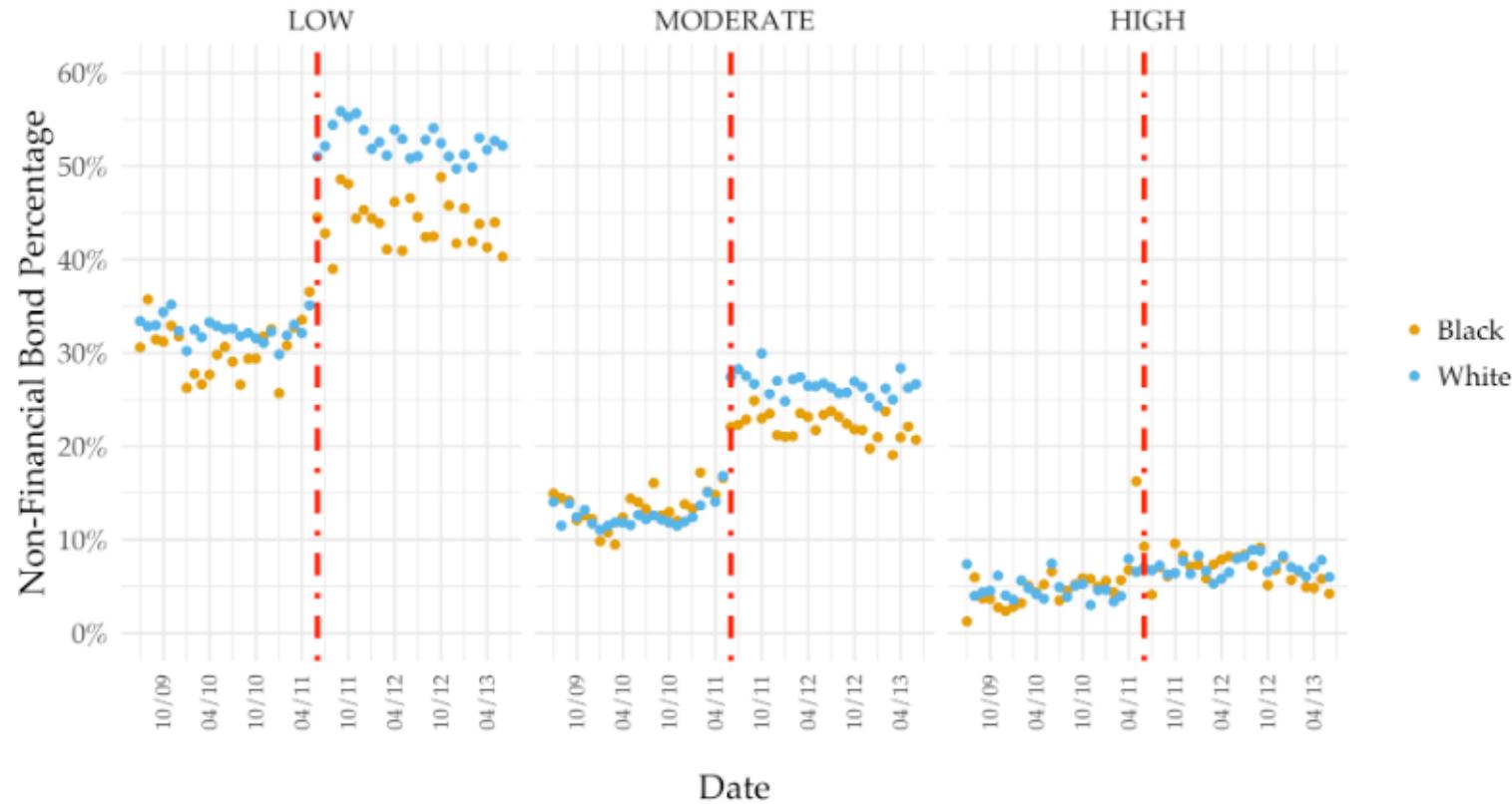
1. Work in small groups – 5 min segments
2. Code all together live

# COMPAS Follow-up

- ▶ Two-year cutoff implementation is wrong
- ▶ Question 19 is highly subjective
- ▶ Thresholds for police searches may be different by groups
- ▶ Judges use risk scores as one input but have final say



Alex Albright, If You Give a Judge a Risk Score, 2019.



Data from Kentucky AOC 7/1/09-6/30/13  
Binned by month-year; red line marks the effective month of HB463

Alex Albright, If You Give a Judge a Risk Score, 2019.

1. COMPAS analysis
2. **What is fairness in machine learning?**
3. Quantitative definitions of fairness in supervised learning
4. Practical tools for analyzing bias
5. Solutions, ethics, and other curveballs

# What is NOT bias in machine learning?

- ▶ **It is not necessarily malicious.**
  - ▶ Bias can occur even when everyone, from the data collectors to the engineers to the medical professionals, have the best intentions.
- ▶ **It is not one and done.**
  - ▶ Just because an algorithm has no bias now does not mean it has no potential later.
- ▶ **It is not new.**
  - ▶ Researchers have raised concerns over the last 50 years.

# What IS bias in machine learning?

- ▶ It is defined many ways, for example **disparate treatment** or **impact of algorithm**. See also, *fairness* or *discrimination*.
- ▶ It is the **culmination of a flawed system**.
  - ▶ Sources including bias in the data collection, bias in the algorithmic process, and bias in the deployment.
- ▶ It is the **vigilance** of how technology can amplify or create bias.

# What are protected classes?

- ▶ Race
- ▶ Sex
- ▶ Religion
- ▶ National origin
- ▶ Citizenship
- ▶ Pregnancy
- ▶ Disability status
- ▶ Genetic information

# Regulated Domains

- ▶ Credit (Equal Credit Opportunity Act)
- ▶ Education (Civil Rights Act of 1964; Education Amendments of 1972)
- ▶ Employment (Civil Rights Act of 1964)
- ▶ Housing (Fair Housing Act)

1. COMPAS analysis
2. What is fairness in machine learning?
3. **Quantitative definitions of fairness  
in supervised learning**
4. Practical tools for analyzing bias
5. Solutions, ethics, and other curveballs

# How do we define “bias”?

- ▶ Fairness through unawareness
- ▶ Group fairness
- ▶ Calibration
- ▶ Error rate balance
- ▶ Representational fairness
- ▶ Counterfactual fairness
- ▶ Individual fairness

# How do we define “bias”?



Arvind Narayanan

@random\_walker

- ▶ Fairness theory

I wrote up a 2-pager titled "21 fairness definitions and their politics" based on the tweetstorm below and it was accepted at a tutorial for the Conference on Fairness, Accountability, and Transparency!

Here it is (with minor edits):

[docs.google.com/document/d/1bn...](https://docs.google.com/document/d/1bn...)

See you on Feb 23/24.

- ▶ Error rate kinds

- ▶ Representativeness

- ▶ Counterfactuals

- ▶ Individual fairness



Arvind Narayanan

@random\_walker · Nov 6, 2017

When I tell my computer science colleagues that there are so many fairness definitions, they are often surprised and/or confused. [Thread]  
[twitter.com/random\\_walker/...](https://twitter.com/random_walker/)

[Show this thread](#)

4:24 PM · Jan 8, 2018 · Twitter Web Client

60 Retweets 208 Likes

# Fairness through unawareness

- ▶ **Idea:** Don't record protected attributes, and don't use them in your algorithm
  - ▶ Predict risk  $Y$  from features  $X$  and group  $A$  using  $P(\hat{Y} = Y|X)$  instead of  $P(\hat{Y} = Y | X, A)$
- ▶ **Pros:** Guaranteed to not be making a judgement on protected attribute
- ▶



# Fairness through unawareness

- ▶ **Idea:** Don't record protected attributes, and don't use them in your algorithm
  - ▶ Predict risk  $Y$  from features  $X$  and group  $A$  using  $P(\hat{Y} = Y|X)$  instead of  $P(\hat{Y} = Y | X, A)$
- ▶ **Pros:** Guaranteed to not be making a judgement on protected attribute
- ▶ **Cons:** Other proxies may still be included in a “race-blind” setting, e.g. zip code or conditions



# Group Fairness

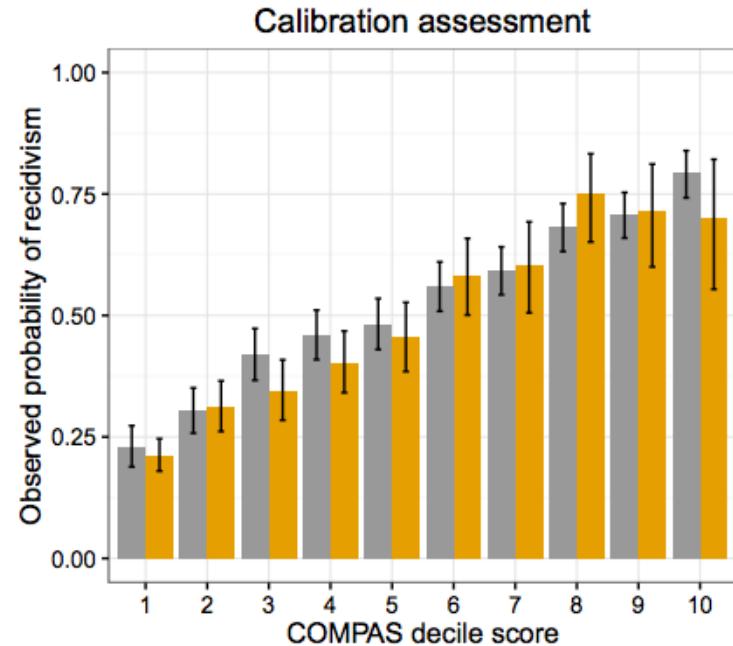
- ▶ **Idea:** Require prediction rate be the same across protected groups
  - ▶ E.g. “20% of the resources should go to the group that has 20% of population”
- ▶ Predict risk  $Y$  from features  $X$  and group  $A$  such that
$$P(\hat{Y} = 1 | A = 1) = P(\hat{Y} = 1 | A = 0)$$
- ▶ **Pros:** Literally treats each race equally
- ▶ **Cons:**
  - ▶
  - ▶

# Group Fairness

- ▶ **Idea:** Require prediction rate be the same across protected groups
  - ▶ E.g. “20% of the resources should go to the group that has 20% of population”
- ▶ Predict risk  $Y$  from features  $X$  and group  $A$  such that
$$P(\hat{Y} = 1 | A = 1) = P(\hat{Y} = 1 | A = 0)$$
- ▶ **Pros:** Literally treats each race equally
- ▶ **Cons:**
  - ▶ Too strong: Groups might have different base rates. Then, even a perfect classifier wouldn't qualify as “fair”
  - ▶ Too weak: Doesn't control error rate. Could be perfectly biased (correct for  $A=0$  and wrong for  $A=1$ ) and still satisfy.

# Calibration

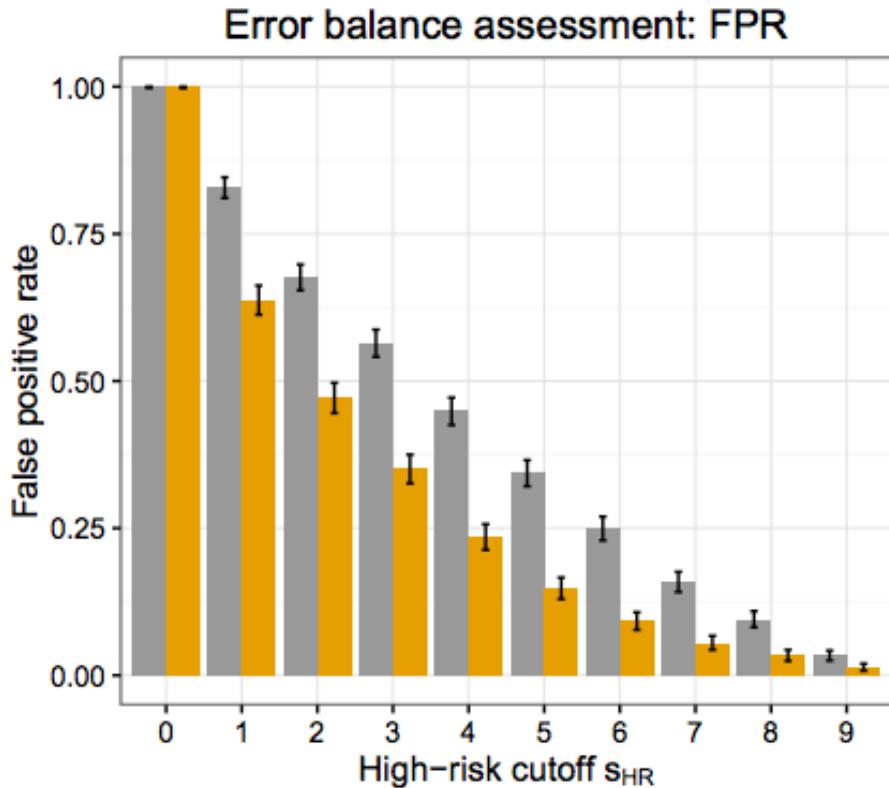
- ▶ **Idea:** Same positive predictive value across groups
  - ▶ Predict  $Y$  from features  $X$  and group  $A$  with score  $S$ :  $P(Y = 1|S = s, A = 1) = P(Y = 1|S = s, A = 0)$
- ▶ **Pros:** “Equally right across groups”
- ▶ **Cons:** Not compatible with error rate balance (next slide)



- ▶ Chouldechova, “Fair prediction with disparate impact”, 2017.

# Error rate balance

- ▶ **Idea:** Equal false positive rates (FPR) across groups
  - ▶  $P(\hat{Y} = 1 | Y = 0, A = 1) = P(\hat{Y} = 1 | Y = 0, A = 0)$
- ▶ **Pros:** “Equally wrong across groups”
- ▶ **Cons:** Incompatible with calibration and false negative rates (FNR), could dilute with easy cases



# Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg \*

Sendhil Mullainathan †

Manish Raghavan ‡

## Abstract

Recent discussion in the public sphere about algorithmic classification has involved tension between competing notions of what it means for a probabilistic classification to be fair to different groups. We formalize three fairness conditions that lie at the heart of these debates, and we prove that except in highly constrained special cases, there is no method that can satisfy these three conditions simultaneously. Moreover, even satisfying all three conditions approximately requires that the data lie in an approximate version of one of the constrained special cases identified by our theorem. These results suggest some of the ways in which key notions of fairness are incompatible with each other, and hence provide a framework for thinking about the trade-offs between them.

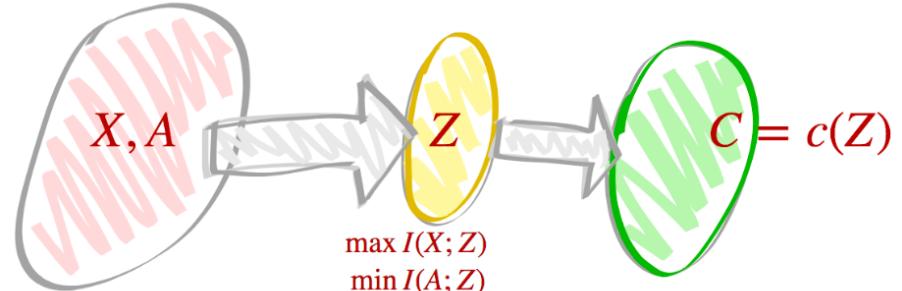
## Inherent Trade-Offs in the Fair Determination of Risk Scores

“We prove that except in highly constrained special cases, **there is no method** that satisfies these three [fairness] conditions simultaneously.”

version of one of the constrained special cases identified by our theorem. These results suggest some of the ways in which key notions of fairness are incompatible with each other, and hence provide a framework for thinking about the trade-offs between them.

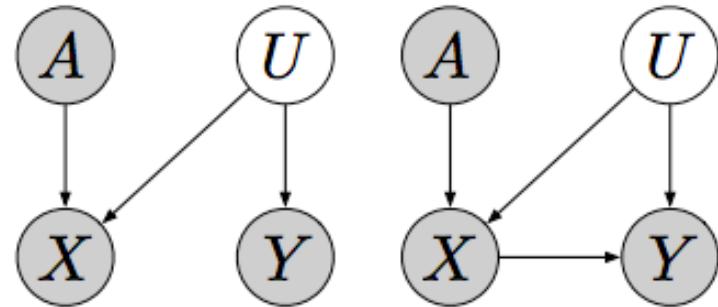
# Representational Fairness

- ▶ **Idea:** Learn latent representation  $Z$  to minimize group information
- ▶ **Pros:** Reduce information given to model but still keep important info
- ▶ **Cons:** Trade-off between accuracy and fairness



# Counterfactual Fairness

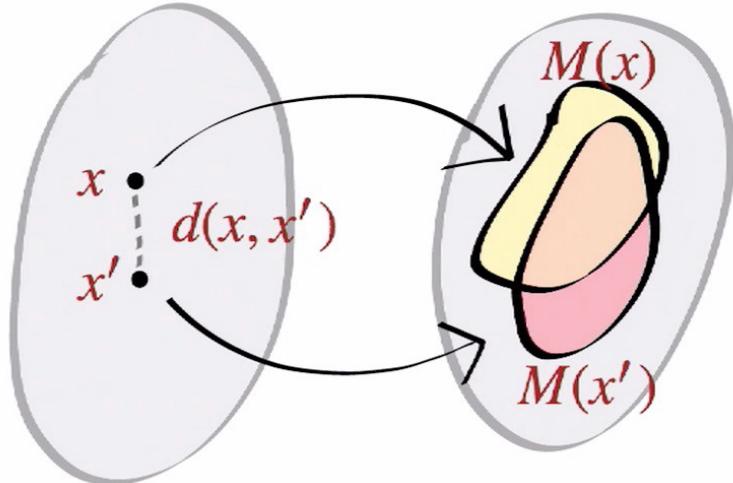
- ▶ **Idea:** Group  $A$  should not cause prediction  $\hat{Y}$
- ▶ **Pros:** Can model explicit connections between variables
- ▶ **Cons:**
  - ▶ Graph model may not actually represent world
  - ▶ Inference assumes observed confounders



$$P(\hat{Y}_{A \leftarrow a}(U) = y | X = x, A = a) \\ = P(\hat{Y}_{A \leftarrow a'}(U) = y | X = x, A = a)$$

# Individual fairness

- ▶ **Idea:** Similar individuals should be treated similarly
- ▶ **Pros:** Can model heterogeneity within each group
- ▶ **Cons:** Notion of “similar” is hard to define mathematically, especially in high dimensions

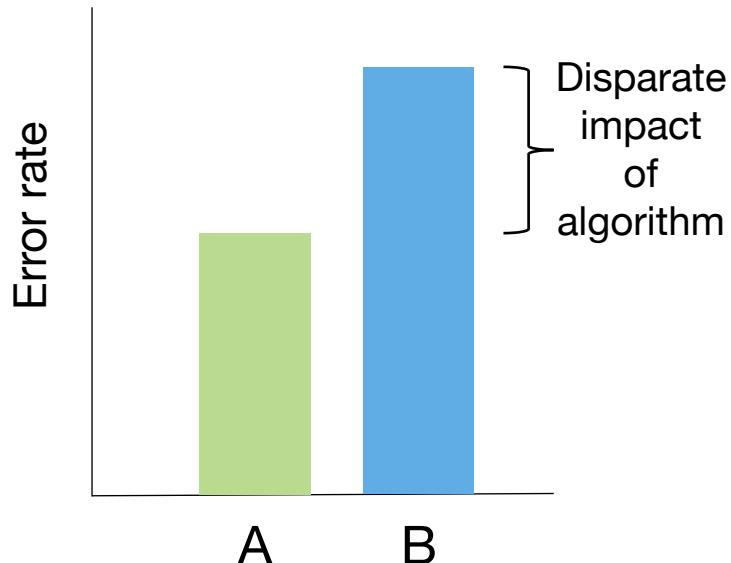


# How do we define “bias”?

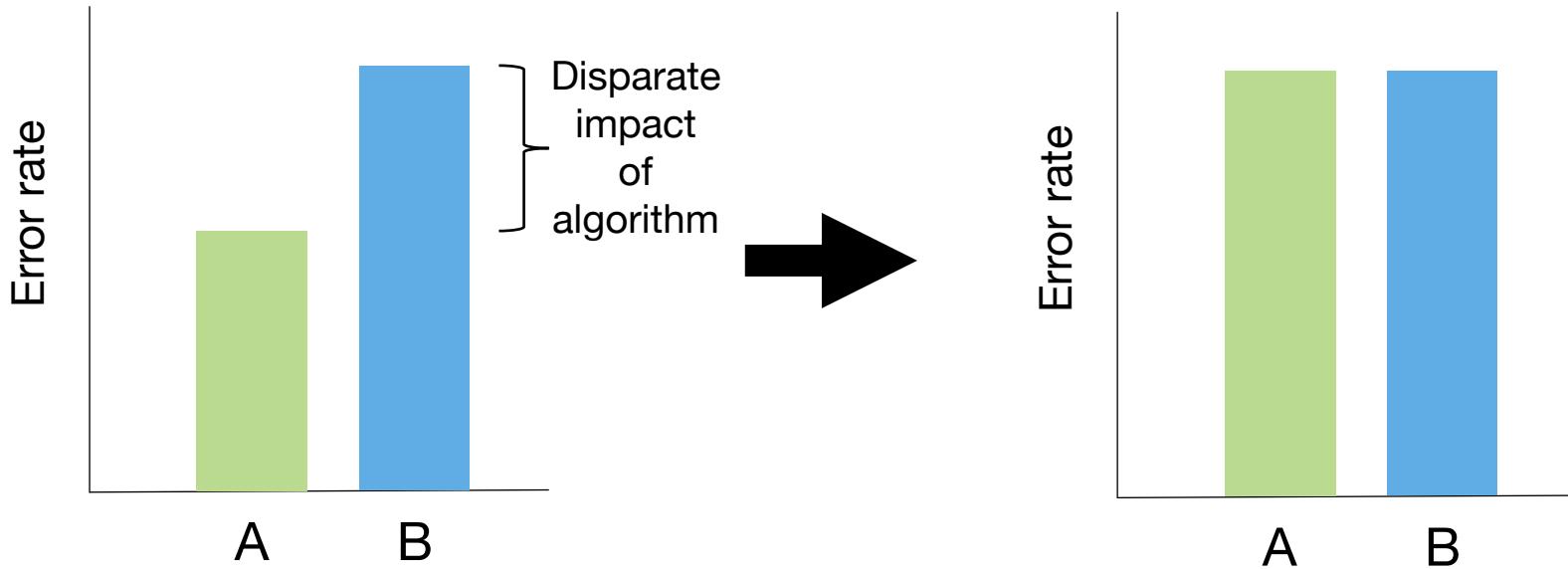
- Fairness through unawareness      Not useful
- **Group fairness**
- **Calibration**      More standard
- **Error rate balance**
- *Representational fairness*      More experimental
- *Counterfactual fairness*
- *Individual fairness*

1. COMPAS analysis
2. What is fairness in machine learning?
3. Quantitative definitions of fairness in supervised learning
4. **Practical tools for analyzing bias**
5. Solutions, ethics, and other curveballs

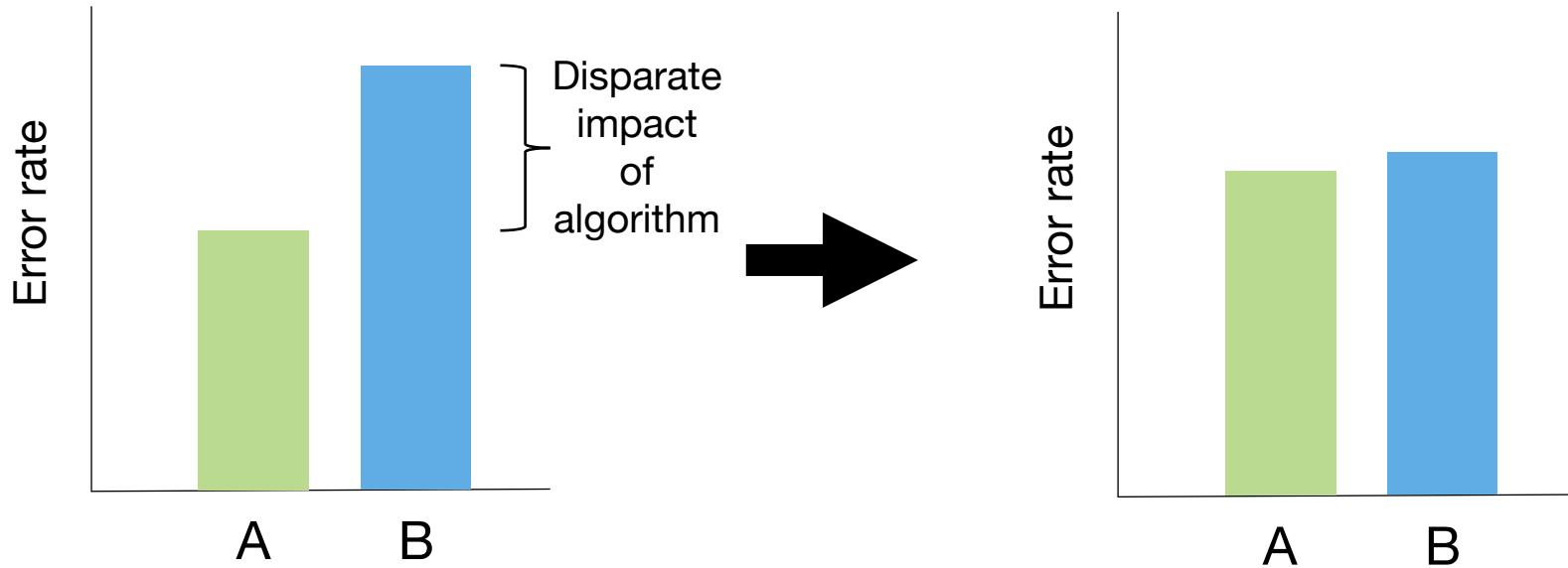
# Tradeoff between accuracy and fairness



# Tradeoff between accuracy and fairness

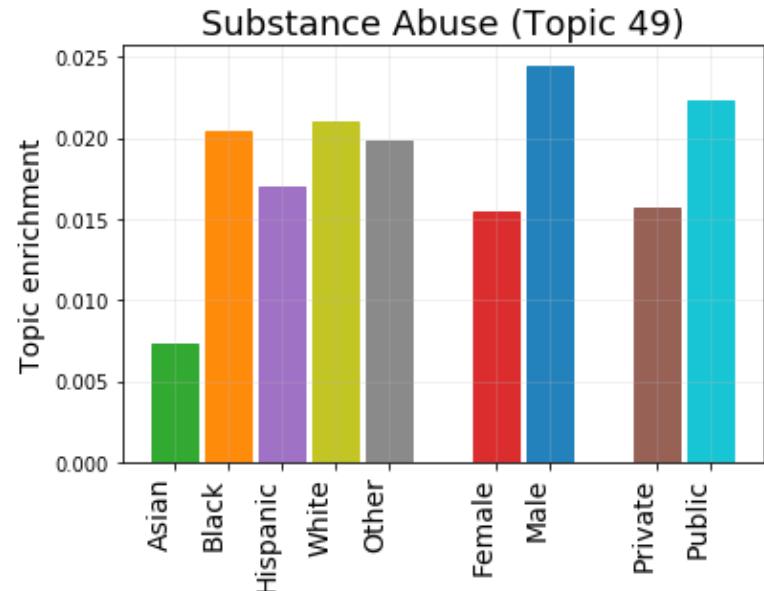


# Tradeoff between accuracy and fairness



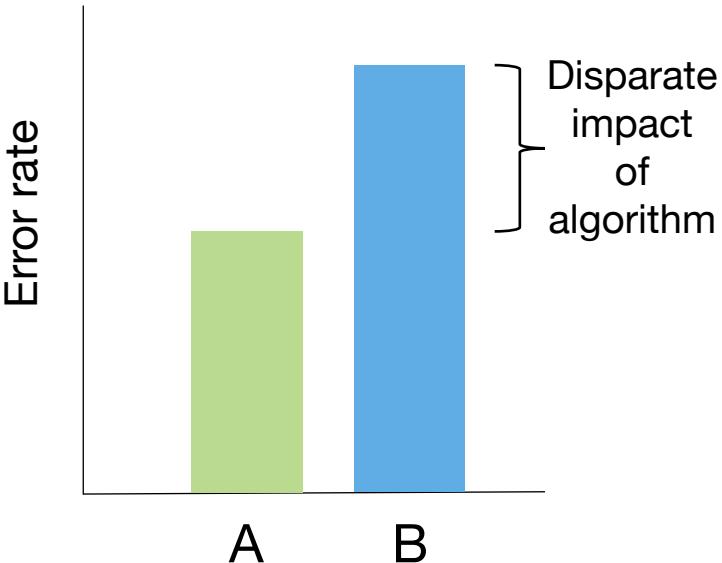
# Understanding data heterogeneity

- ▶ We can understand **unstructured psychiatric notes** through LDA topic modeling
- ▶ One salient topic, **substance abuse**, had the following key words: use, substance, abuse cocaine, mood, disorder, dependence, positive, withdrawal, last, reports, ago, day, drug



# Consider bias, variance, noise

	Description
<b>Bias</b>	How well the model fits the data
<b>Variance</b>	How much the sample size affects the accuracy
<b>Noise</b>	Irreducible error independent of sample size and model



## RESEARCH ARTICLE

## ECONOMICS

# Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer<sup>1,2\*</sup>, Brian Powers<sup>3</sup>, Christine Vogeli<sup>4</sup>, Sendhil Mullainathan<sup>5\*†</sup>

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses.

Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

## RESEARCH ARTICLE

“The bias arises because the algorithm **predicts health care costs rather than illness** … despite health care cost appearing to be an effective **proxy for health**”

Help from AI to reduce the bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

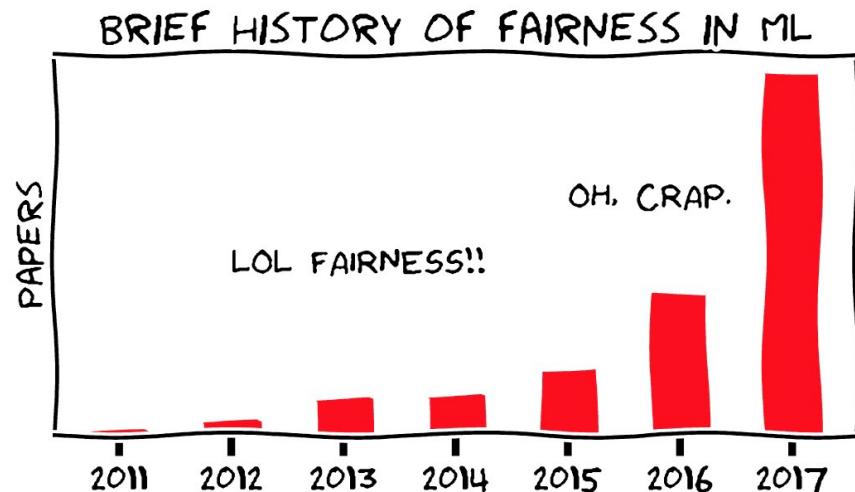
1. COMPAS analysis
2. What is fairness in machine learning?
3. Quantitative definitions of fairness in supervised learning
4. Practical tools for analyzing bias
5. **Solutions, ethics, and other curveballs**



AI NOW  
INSTITUTE

# Fair ML for Health

NeurIPS 2019 Workshop, East Ballroom B, Vancouver, Canada



# Inclusive ML guide - AutoML

At Google, we've been thinking hard about the [principles](#) that motivate and shape our work in artificial intelligence (AI). We're committed to a human-centered approach that foregrounds [responsible AI practices](#) and products that work well for all people and contexts. These values of responsible and inclusive AI are at the core of the AutoML suite of machine learning products, and manifest in the following ways.

AI

IBM Research Releases ‘Diversity in Faces’ Dataset to Advance Study of Fairness in Facial Recognition Systems

NO **TECH** for



**PLEASE DON'T**



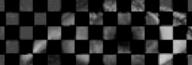
**INCLUDE US**

**WORKSHOP**

**APPLY NOW**

**[digitaljusticelab.ca/cfp](http://digitaljusticelab.ca/cfp)**

Rolling Applications until Sept 4th 2019



ECONOMIC VIEW

## *Biased Algorithms Are Easier to Fix Than Biased People*

Racial discrimination by algorithms or by people is harmful — but that's where the similarities end.

By Sendhil Mullainathan

Dec. 6, 2019



# Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination

Marianne Bertrand, Sendhil Mullainathan

NBER Working Paper No. 9873

Issued in July 2003

NBER Program(s):Labor Studies Program

We perform a field experiment to measure racial discrimination in the labor market. We respond with fictitious resumes to help-wanted ads in Boston and Chicago newspapers. To manipulate perception of race, each resume is assigned either a very African American sounding name or a very White sounding name. The results show significant discrimination against African-American names: White names receive 50 percent more callbacks for interviews. We also find that race affects the benefits of a better resume. For White names, a higher quality resume elicits 30 percent more callbacks whereas for African Americans, it elicits a far smaller increase. Applicants living in better neighborhoods receive more callbacks but, interestingly, this effect does not differ by race. The amount of discrimination is uniform across occupations and industries. Federal contractors and employers who list Equal Opportunity Employer<sup>1</sup> in their ad discriminate as much as other employers. We find little evidence that our results are driven by employers inferring something other than race, such as social class, from the names. These results suggest that racial discrimination is still a prominent feature of the labor market.

# Open questions

- ▶ How can we build inclusive algorithms and datasets?
- ▶ For what settings should we use algorithms?
- ▶ Can we ever promise an algorithm is “fair”?
- ▶ When should we use humans and when should we use algorithms?

# Looking forward

- ▶ Researchers have made great progress **auditing bias** in existing wide-spread algorithms.
- ▶ **Formalizing fairness** quantitatively can build fairness constraints directly into high-stakes models.
- ▶ Long-term solutions include **growing research community, rethinking datasets, and considering societal impacts.**