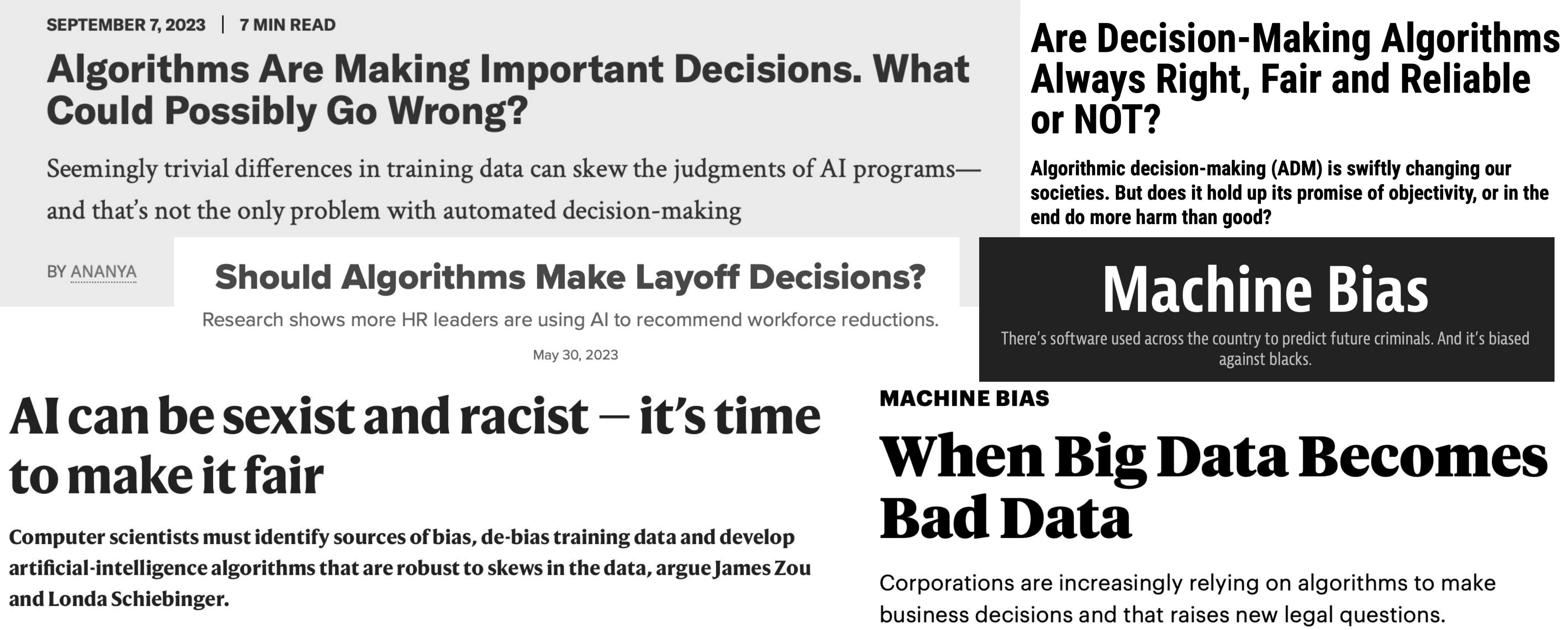
# Algorithmic Bias and Statistical Notions of Fairness

Dasha Asienga

Advisor: Professor Katharine Correia

Fall Thesis Talk

# Data-Driven Algorithms Can Perpetuate, or even Amplify, Existing Human Biases



SEPTEMBER 7, 2023 | 7 MIN READ

## Algorithms Are Making Important Decisions. What Could Possibly Go Wrong?

Seemingly trivial differences in training data can skew the judgments of AI programs—and that's not the only problem with automated decision-making

BY ANANYA

## Are Decision-Making Algorithms Always Right, Fair and Reliable or NOT?

**Algorithmic decision-making (ADM) is swiftly changing our societies. But does it hold up its promise of objectivity, or in the end do more harm than good?**

## Should Algorithms Make Layoff Decisions?

Research shows more HR leaders are using AI to recommend workforce reductions.

May 30, 2023

## Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

## AI can be sexist and racist — it's time to make it fair

**Computer scientists must identify sources of bias, de-bias training data and develop artificial-intelligence algorithms that are robust to skews in the data, argue James Zou and Londa Schiebinger.**

MACHINE BIAS

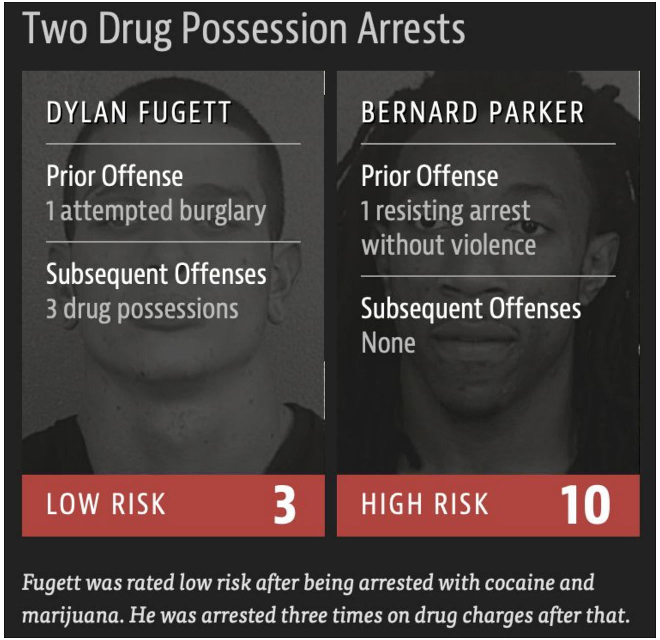
## When Big Data Becomes Bad Data

Corporations are increasingly relying on algorithms to make business decisions and that raises new legal questions.

**Algorithmic bias** arises when an algorithm’s decisions are skewed towards a particular group of people, either positively or negatively.

# Examples of Algorithmic Bias

## Criminal Justice (COMPAS):

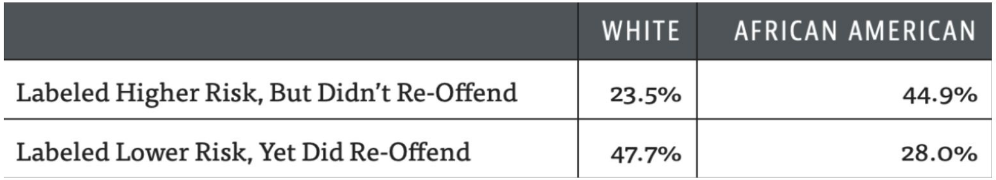


Two Drug Possession Arrests

| DYLAN FUGETT | BERNARD PARKER |
|---|---|
| Prior Offense<br>1 attempted burglary | Prior Offense<br>1 resisting arrest without violence |
| Subsequent Offenses<br>3 drug possessions | Subsequent Offenses<br>None |
| LOW RISK 3 | 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.*

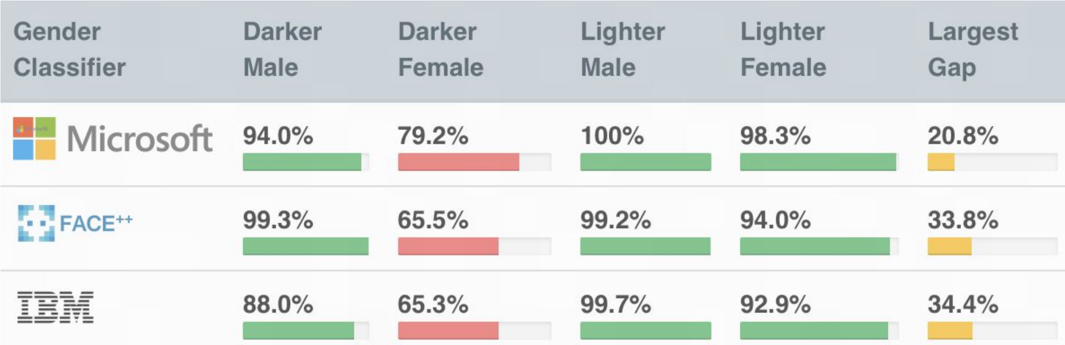
Equal accuracy for both African-American offenders and White offenders BUT higher false positive rates for African-American offenders, and vice versa.



| | 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% |

Figure 1.1: Prediction Fails Differently for Black v White Defendants

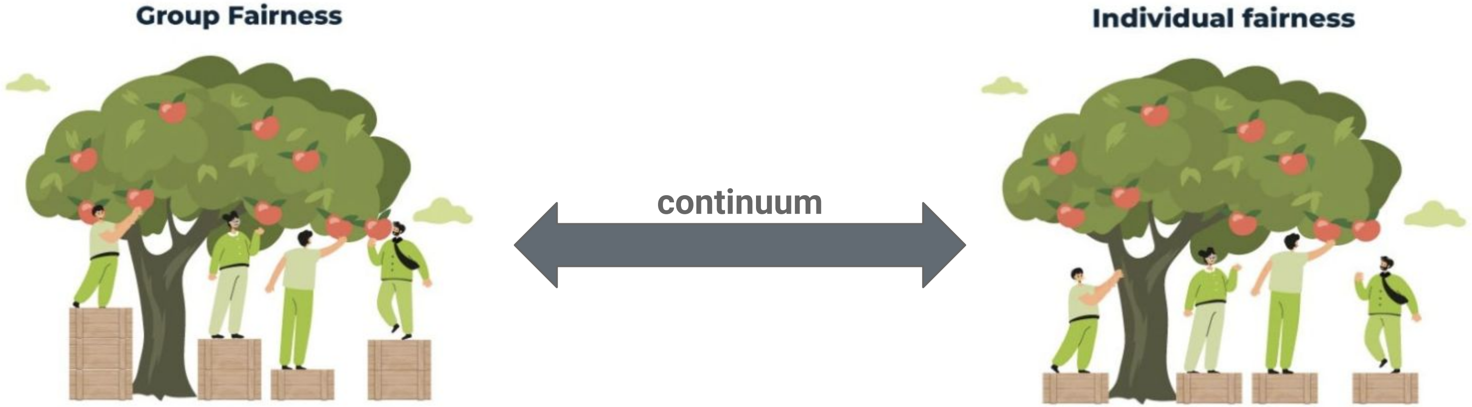
## Facial Recognition:



| 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% |

# How can we mathematically/ statistically measure a model’s (un)fairness?

# Group v Individual Statistical Definitions of Fairness



Fix a few demographic groups and assess the **parity of some statistical measures across all the groups**.

*Does not guarantee fairness to individuals or structured subgroups.
*Focuses on "average numbers".

**Similar individuals should be treated similarly** along some defined similarity or inverse distance metrics.

*Can be impractical, relies on strong assumptions about the data, and approaches the realm of causality.

# Notation

$Y \longrightarrow$ Response variable: real-valued/ continuous (for regression) or discrete (for classification).

$\hat{Y} \longrightarrow$ Predicted response value.

$A \longrightarrow$ The set of demographic groups: the **sensitive/ protected group** is at a disadvantage.

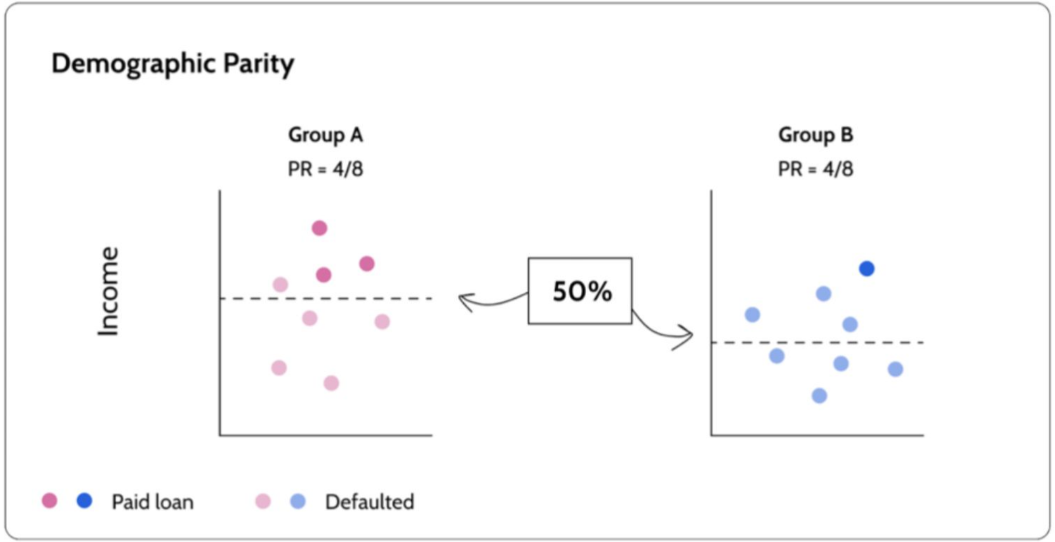
$R \longrightarrow$ The set of all possible ratings (a covariate in the data set); for example, a qualification rating for participants in a hiring algorithm.

$X \longrightarrow$ The set of all predictor variables.

# Independence

$$P(\hat{Y}=1|A=a)=P(\hat{Y}=1|A=b),\ \forall a,b\in A,$$

where a, b are the two demographic groups in question.



- Requires $\hat{Y} \perp\!\!\!\perp A$
- Also known as demographic parity or statistical parity.
- The likelihood of a positive outcome should be the same across each demographic group.

# Conditional Demographic Parity

$$P(\hat{Y}=1|A=a,R=r)=P(\hat{Y}=1|A=b,R=r),\ \forall a,b\in A,\forall r.$$

i.e. $\hat{Y} \perp\!\!\!\perp A|R$ with R as the set of possible ratings.

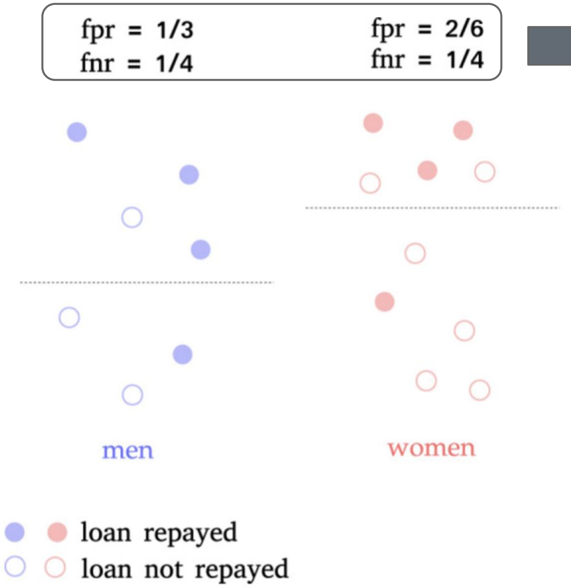
**Individual Fairness**

$$P(\hat{Y}=1|A=a,X=x)=P(\hat{Y}=1|A=b,X=x),\ \forall a,b\in A,\forall x\in X.$$

i.e. $\hat{Y} \perp\!\!\!\perp A|X$ with X as the set of all explanatory variables.

# Separation

$$P(\hat{Y} = 1 | A = a, Y = y) = P(\hat{Y} = 1 | A = b, Y = y),\ \forall a, b \in A,\ \ y \in \{0, 1\}$$



$$FPR = \frac{FP}{FP + TN}$$

$$FNR = \frac{FN}{TP + FN}$$

| | | True Class | |
|---|---|---|---|
| | | Positive | Negative |
| Predicted Class | Positive | TP | FP |
| | Negative | FN | TN |

- Requires $\hat{Y} \perp\!\!\!\perp A | Y$ .
- Also known as the equality of odds.
- The error rates should be the same across each demographic group.

# Sufficiency

$$P(Y = 1|A = a, \hat{Y} = 1) = P(Y = 1|A = b, \hat{Y} = 1),\ \forall a, b \in A.$$

- Requires $Y \perp\!\!\!\perp A|\hat{Y}$.
- Also known as predictive parity.
- The precision of the model should be equal across all demographic groups.

We can continue to define more fairness metrics, but this begs the fundamental question: **can they be enforced simultaneously**?

# Fairness Conflict – Classification

**sufficiency:** [equal positive predictive values (PPV)]

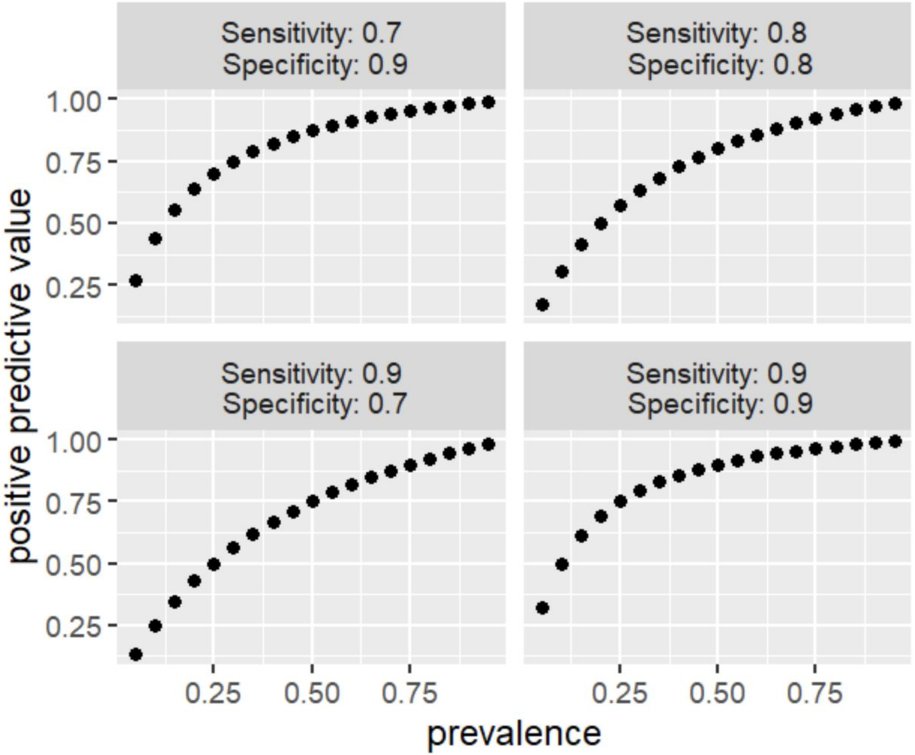
**separation:** [equal FPR and FNR]

Define:

➔ FNR = 1 – *sensitivity*
➔ FPR = 1 – *specificity*

Then, given values of PPV ∈ (0, 1) and prevalence *p* ∈ (0, 1), we can show that:

$$\mathrm{FPR} = \frac{p}{1-p}\,\frac{1-\mathrm{PPV}}{\mathrm{PPV}}(1-\mathrm{FNR}).$$



# Fairness Conflict – Regression

Consider a case with **gender *G*** as the protected attribute.

These two cannot hold simultaneously
**if the average distribution of Y is different for both groups**.

A model that satisfies independence:

$$E[\hat{Y}|G = Male] = E[\hat{Y}|G = Female]$$

cannot simultaneously satisfy
equal error rates:

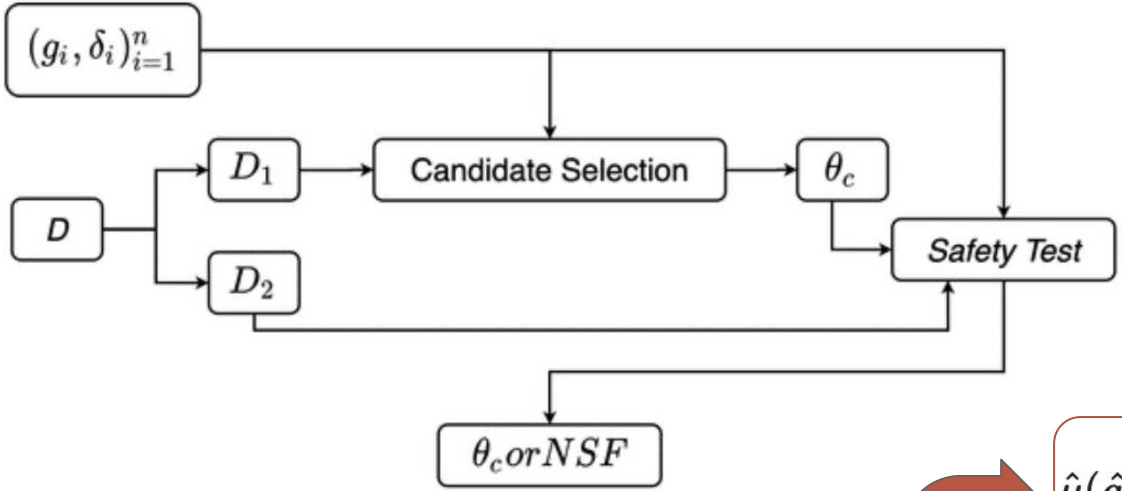
$$E[\hat{Y} - Y|G = Male] = E[\hat{Y} - Y|G = Female]$$

# What can we do in the face of this conflict?

# The Seldonian Algorithm

The algorithm is premised on the notion that if **‘unfair’** or **‘unsafe’** outcomes or behaviors can be **defined mathematically**, then it should be possible to create algorithms that can **learn from the data on how to avoid these unwanted results** with **high confidence**.

# An Overview of the Seldonian Framework



**Results from a toy regression example:**
- Better performance with **more data**.
- Probability of a solution stabilized at 0.8 as the amount of data increased.
- There is an **accuracy tradeoff**.
- The solution returned almost always satisfied the constraint with 100% confidence.

$$\hat{\mu}(\hat{g}_i(\theta_c, D_1)) + 2\frac{\hat{\sigma}(\hat{g}_i(\theta_c, D_1))}{\sqrt{|D_2|}} t_{1-\delta_i,|D_2|-1} \leq 0$$

**The Seldonian Optimization Problem:**

$$\arg\max_{a \in \mathscr{A}} f(a) \text{ s.t. } \forall i \in \{1, \ldots, n\}, \Pr(g_i(a(D)) \leq 0) \geq 1 - \delta_i$$

# Next Steps

1. Complete simulations and experimentation on the Seldonian Algorithm and further understand fairness conflicts

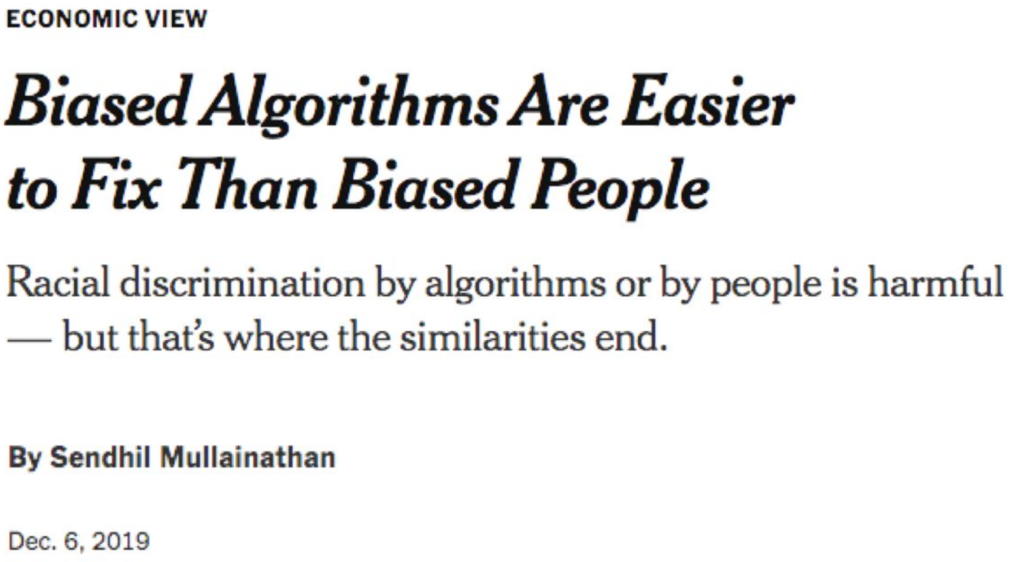
2. Fit Seldonian Algorithm on GPA data set from a Brazilian university (regression)

3. Fit Seldonian Algorithm on COMPAS data set (classification)

# Thank You!

Questions?

# Good News!



ECONOMIC VIEW

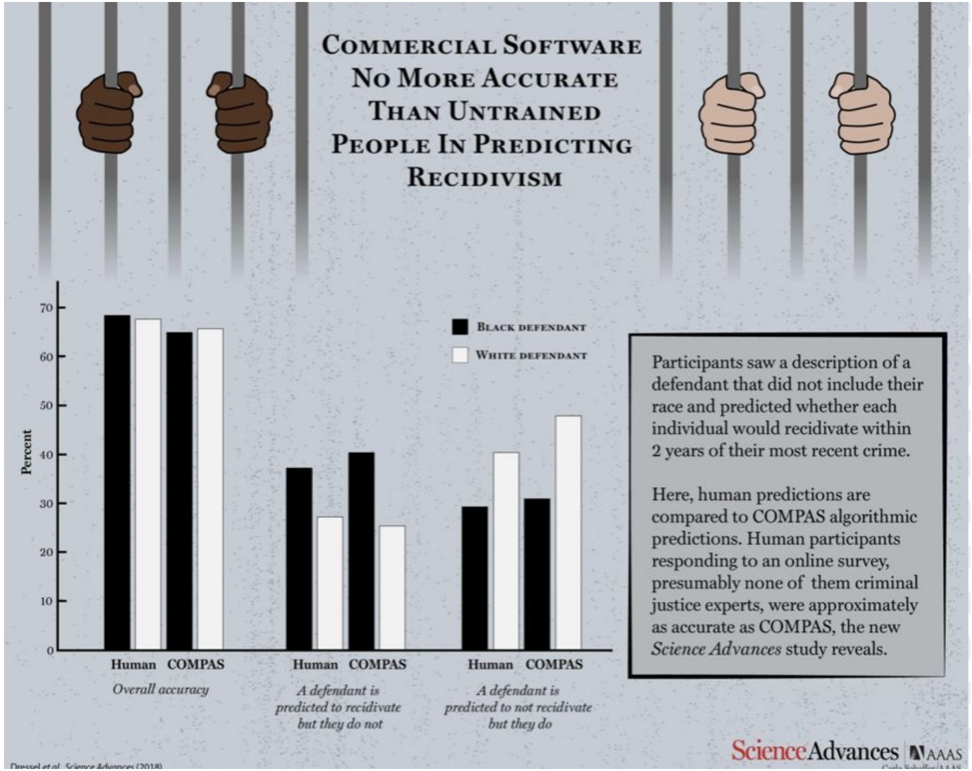
## *Biased Algorithms Are Easier to Fix Than Biased People*

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

By Sendhil Mullainathan

Dec. 6, 2019

# COMPAS: Correctional Offender Management Profiling for Alternative Sanctions



Participants saw a description of a defendant that did not include their race and predicted whether each individual would recidivate within 2 years of their most recent crime.

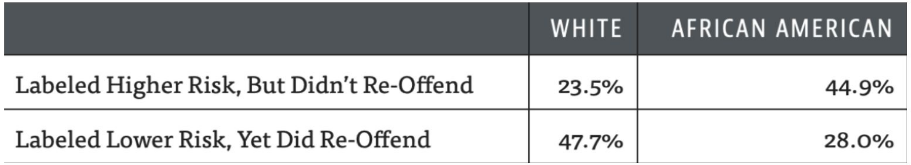
Here, human predictions are compared to COMPAS algorithmic predictions. Human participants responding to an online survey, presumably none of them criminal justice experts, were approximately as accurate as COMPAS, the new *Science Advances* study reveals.

**Purpose:** predict a defendant's risk of recidivism – committing another crime – within two years (the data set has 7000 observations).

**Bias:** Higher false positive rates for African-American offenders than Caucasian offenders.

**Danger:** Across the country, scores of similar assessments are given to judges, which injects bias into courts.

**The overall accuracy is 61%, BUT**

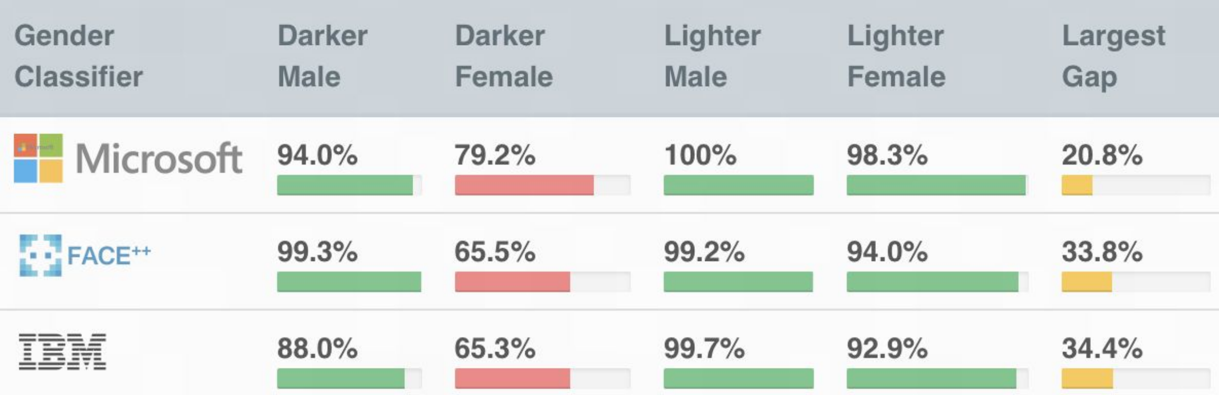


| | WHITE | AFRICAN AMERICAN |
|---|---|---|
| Labeled Higher Risk, But Didn't Re-Offend | 23.5% | 44.9% |
| Labeled Lower Risk, Yet Did Re-Offend | 47.7% | 28.0% |

Figure 1.1: Prediction Fails Differently for Black v White Defendants

# What is Algorithmic Bias?

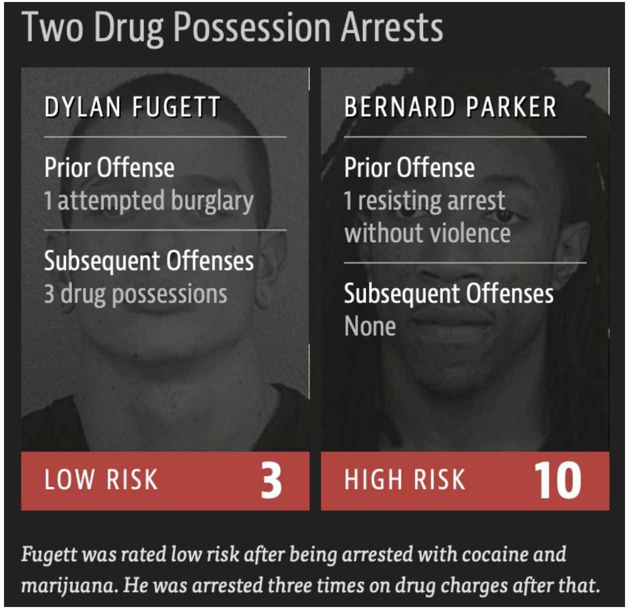
**Facial Recognition:**



| 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% |

**Algorithmic bias** arises when an algorithm's decisions are skewed towards a particular group of people, either positively or negatively.

**Criminal Justice:**



Two Drug Possession Arrests

DYLAN FUGETT

Prior Offense
1 attempted burglary

Subsequent Offenses
3 drug possessions

LOW RISK 3

BERNARD PARKER

Prior Offense
1 resisting arrest without violence

Subsequent Offenses
None

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.*