

# Value-Sensitive Design in Machine Learning: Fairness and Data Subjects

Ethics Lecture for DS4400, Spring 2022

**Vance Ricks**

Associate Teaching Professor  
of Philosophy and Computer  
Science

Department of  
Philosophy/Religion,  
Northeastern University

[v.ricks@northeastern.edu](mailto:v.ricks@northeastern.edu)

# Agenda for the Week

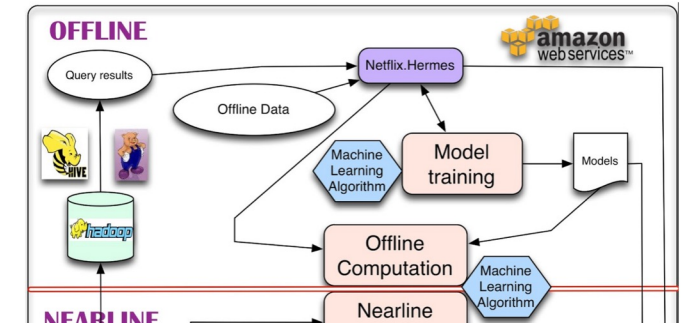
- Today
  - Introduction: machine learning algorithms in the wild
  - SOTBF: using a simulation to uncover ethical questions
  - Three conceptions of “fairness” and “unfairness”
  - KHASM: Treating people as data subjects
  - A short introduction to VSD (value-sensitive design)
- Wednesday
  - Quick review of Monday’s materials
  - Articulating values and identifying stakeholders: using value-sensitive design (VSD)
  - Revisiting SOTBF and KHASM
  - Conclusion: Keeping the human in machine learning

# Today

1. Introduction: machine learning algorithms in the wild
2. SOTBF: using a simulation to (re-)uncover ethical issues in ML
3. Three conceptions of “fairness” and “unfairness”
4. KHASM: Treating people as data subjects
5. A short introduction to VSD (value-sensitive design)

# Machine Learning Based Computer Aided Diagnosis of Breast Cancer Utilizing Anthropometric and Clinical Features

# How To Design A Spam Filtering System with Machine Learning Algorithm



# Machine Learning Algorithms In the Wild



# Machine Learning Algorithms In the Wild

Both Zoom and Twitter found themselves under fire this weekend for their respective issues with algorithmic bias. On Zoom, it's an issue with the video conferencing service's virtual backgrounds and on Twitter, it's an issue with the site's photo cropping tool.

It started when [Ph.D. student Colin Madland tweeted](#) about a Black faculty member's issues with Zoom. According to Madland, whenever said faculty member would use a virtual background, Zoom would remove his head.

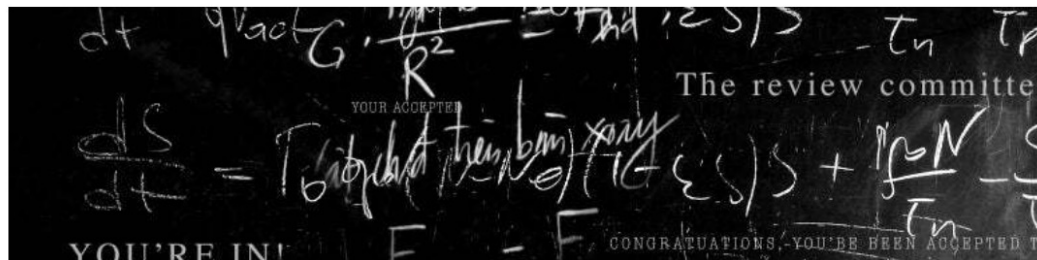
In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to people familiar with the matter. They did not specify the names of the schools.

## Media Coverage

[Can Algorithms Select Students "Most Likely to Succeed"?](#)

*by: Rebecca Koenig*

*July 10, 2020*










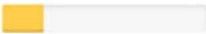





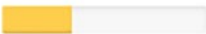





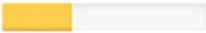
*Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)*

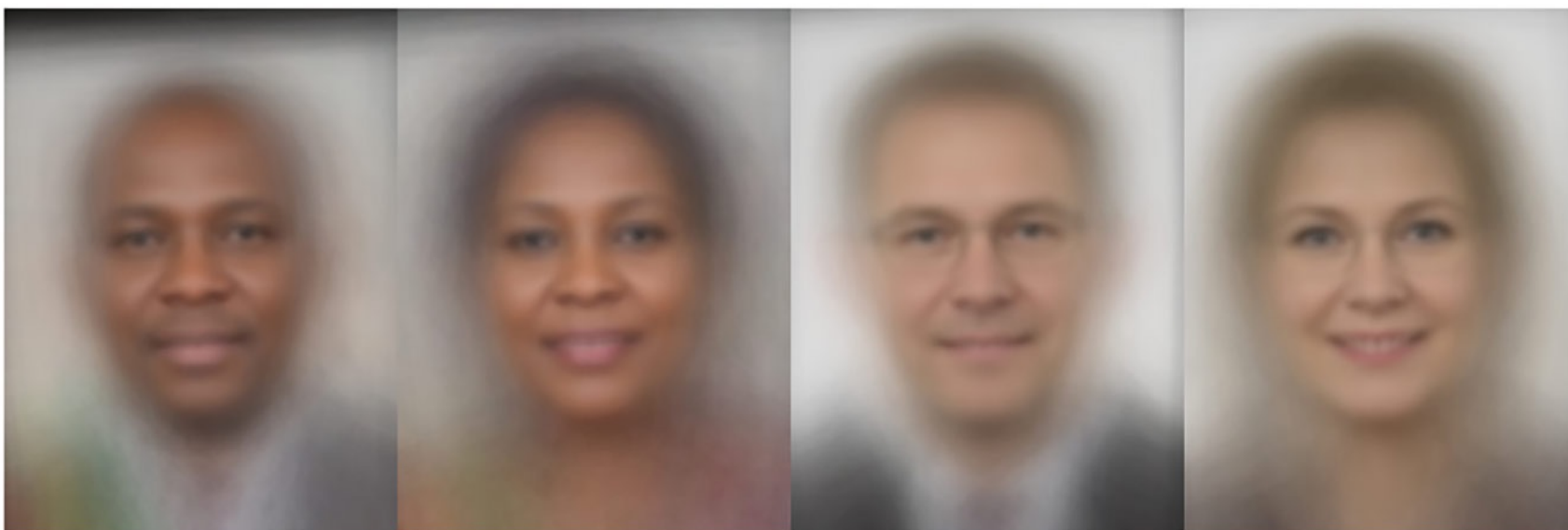
# Machine Bias

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

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

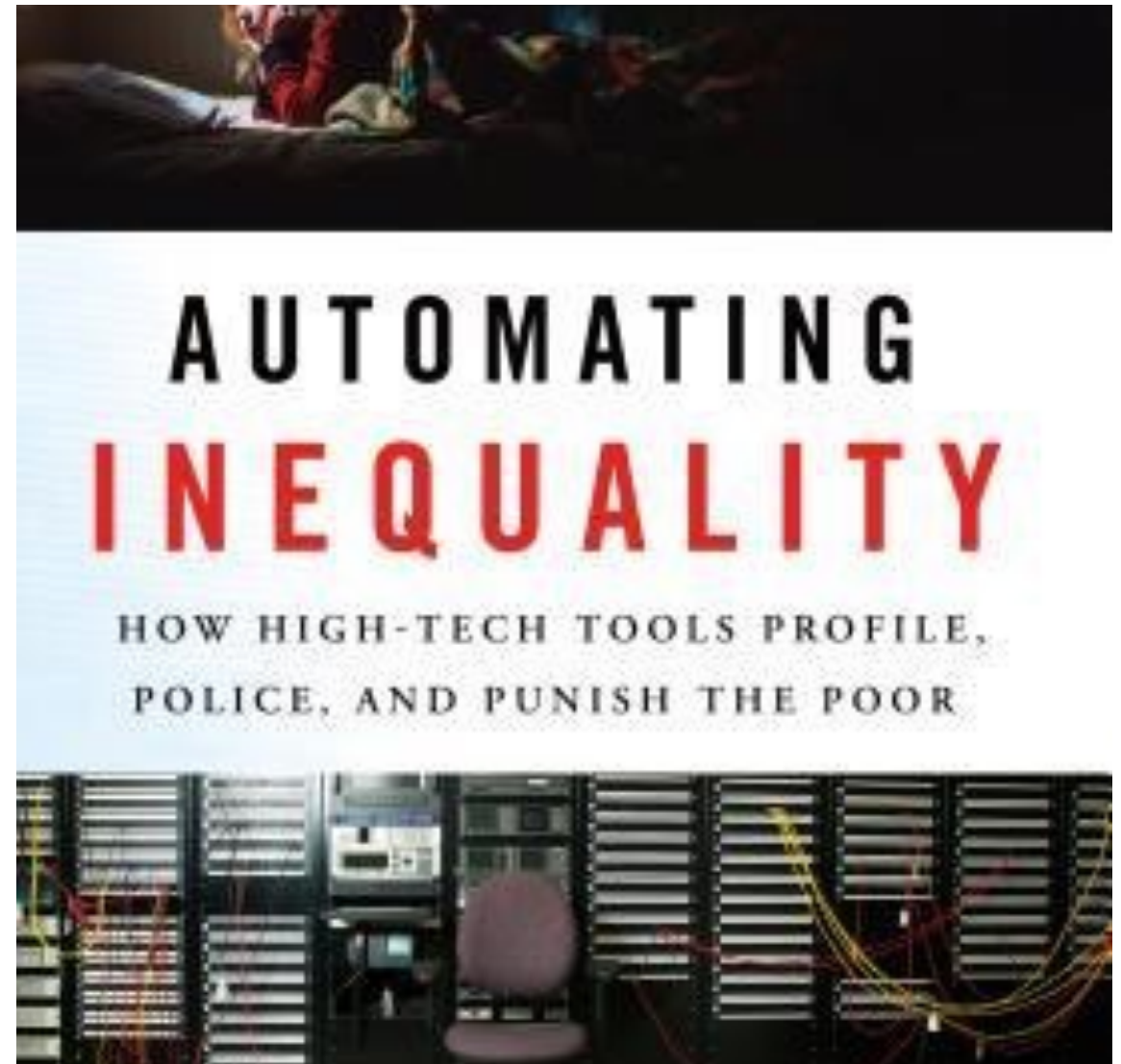




The Family and Social Services Administration (FSSA) of Indiana provides welfare, food stamps, public health insurance

---

- goals defined as to reduce fraud, spending and number of those on welfare
- prior to automation, FSSA erred on side of providing benefits: False Pos rate = 4.4% False Neg rate = 1.5%
- after automation, erred on opposite side: FP rate = 6.2% FN rate = 12.2%
- when denied, no explanation given for why
- did not use records from previous system, requiring all new applications




# Preliminary Questions for Small Group (3 – 5 people) Discussions

## Instructions:

In your group, take about 5 minutes to discuss and answer both of the questions below.

Jot down your answers, to report back to the rest of the class.



**Question One:** What is *fair treatment*, as opposed to unfair treatment?

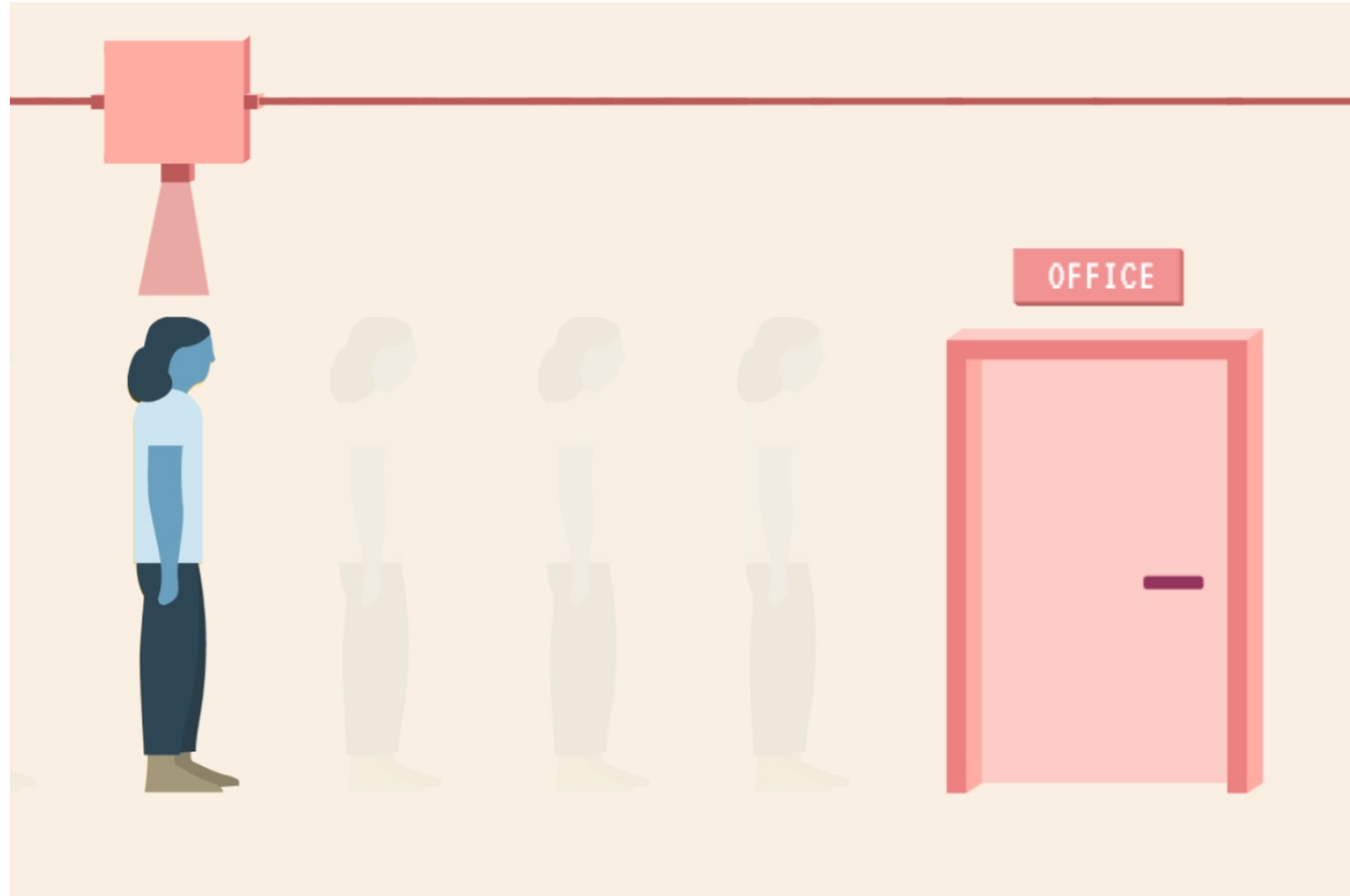


**Question Two:** Can you think of a case where *fair treatment* does not result in a fair *outcome*?

# Collected Group Responses

- Fair vs Unfair treatment
  - Considering only relevant factors in decision (e.g., for hiring)
  - Not being influenced by bias or prejudices
  - Implies public, known rules that people know about
  - Implies some kind of comparison among groups
  - Unfair: Get something you don't deserve (implies a connection to desert)
  - Give everybody what they need
- Different definitions of fairness
  - Demographic parity: Equal representation, given the proportions

# Survival of the Best Fit (SOTBF): (re-)uncovering ethical issues in ML



## Instructions for SOTBF:

1. Go to  
[www.survivalofthebestfit.com](http://www.survivalofthebestfit.com)



2. (Individually,) Play the  
simulation TWICE, as follows:

a. Play it once, aiming to treat each of the applicants “fairly”, **in accordance with whatever your group said counts as “fair treatment”**.

b. Play it a second time, aiming to achieve “a fair outcome”, **in accordance with whatever your group said counts as “a fair outcome”**.



An illustration of a person with dark hair, wearing a light blue shirt and dark pants, standing on the left side of the frame. To their right is a vertical list of five topics, each preceded by a grey circle. Further right is a red door with a sign above it that says 'OFFICE'. The background is a light beige color with a red horizontal line at the top and a white horizontal line at the bottom.

Fairness(es)

Bias(es)

Training data collection practices

“Comprehensibility” of the  
algorithms

People as data subjects

OFFICE

Discussion of SOBTF results

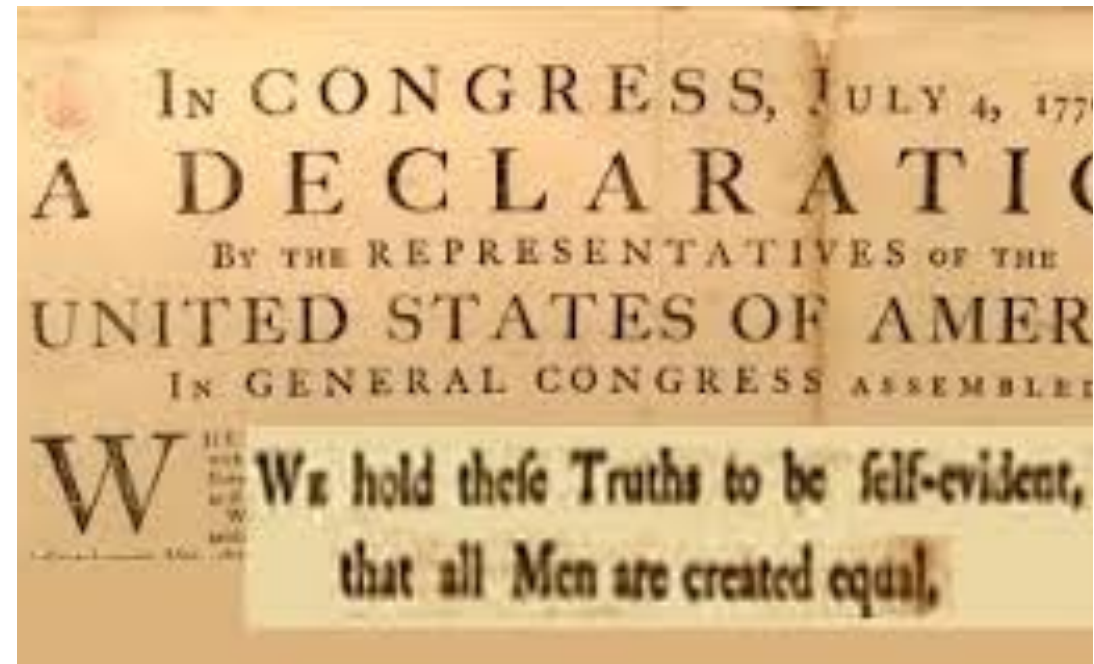
# Three conceptions of “fairness” and “unfairness”

# 1. Fair treatment as a MORAL norm: *people* are treated fairly when those who are similarly situated are treated similarly (when like cases are treated alike).

Any decision to treat *classes* of persons differently should be rationally related to achieving the purpose for which the classification is made. That principle reflects the moral equality of persons.

For example, suppose Prof. Oprea decides to give “A”s to everyone with curly hair and give everyone else “F”s.

This is unfair to **both** groups of students, because your hair texture isn’t relevant to what your grade in this course should be.



## 2. “Fair treatment” as a legal norm

---

“No state shall make or enforce any law which shall abridge the privileges or immunities of citizens of the United States; nor shall any state deprive any person of life, liberty, or property, without due process of law; nor deny to any person within its jurisdiction the **equal protection** of the laws” (14<sup>th</sup> Amendment).



## What is “equal protection of the laws”?

- 1. any [suspect, possibly suspect, not suspect] classification of people
- 2. must be [necessarily, substantially, or merely rationally] related to achieving
- 3. a [compelling, important, or merely legitimate] state purpose

### 3. Fairness as a *distributive* norm:

“patterned” views

v

“process/  
procedural” views

#### Patterned views

Given some goods that we want to distribute,

A **distribution** of those goods is fair if and only if it distributes them in accordance with some (morally acceptable) **property or pattern**

(every *citizen of the US* has the same set of Constitutionally-guaranteed rights)

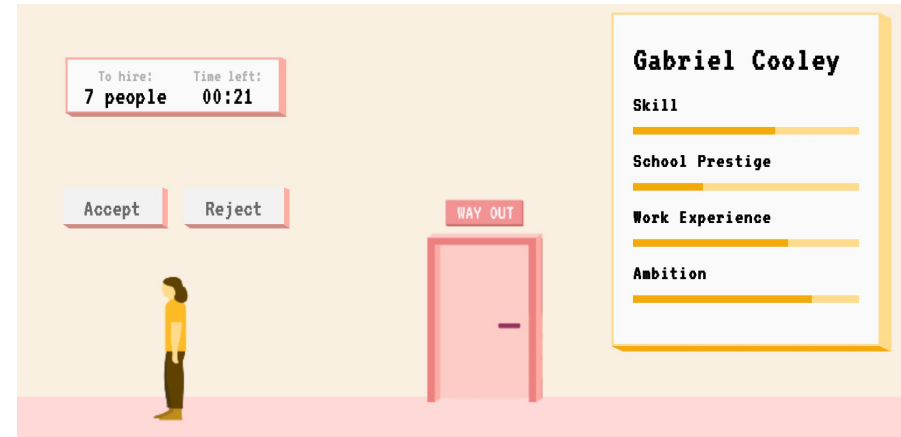
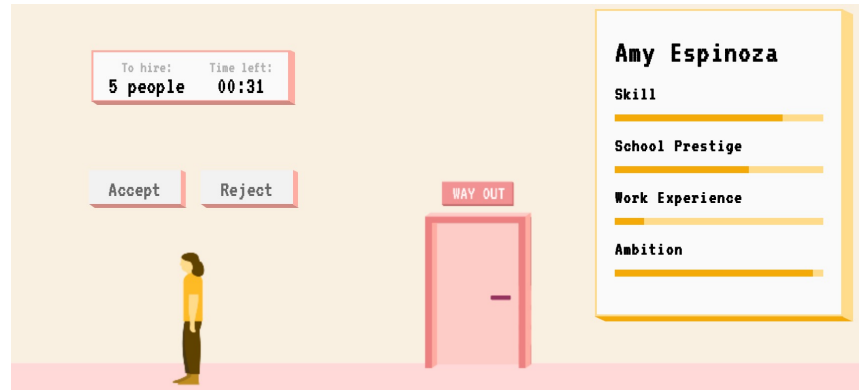
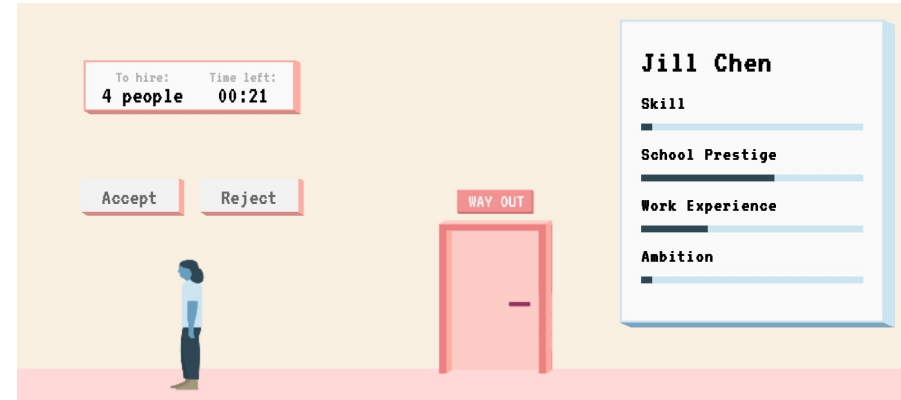
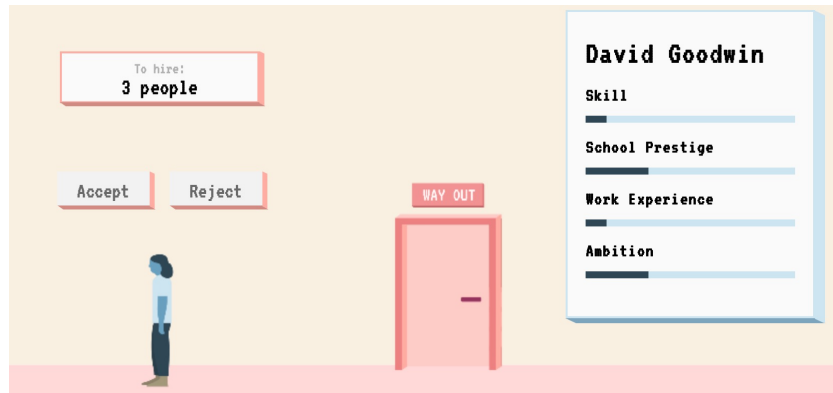
#### Process/Procedural views

Given some goods that we want to distribute,

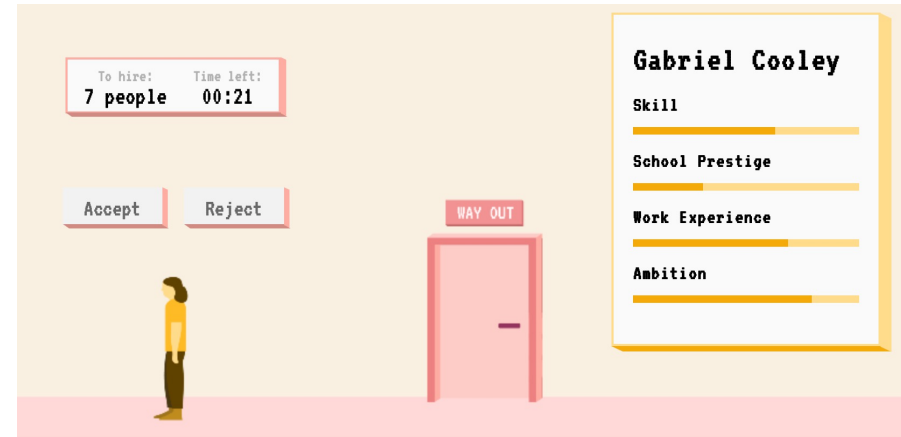
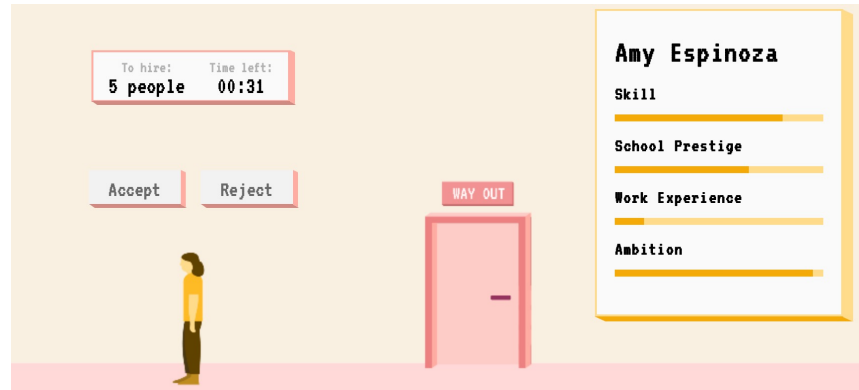
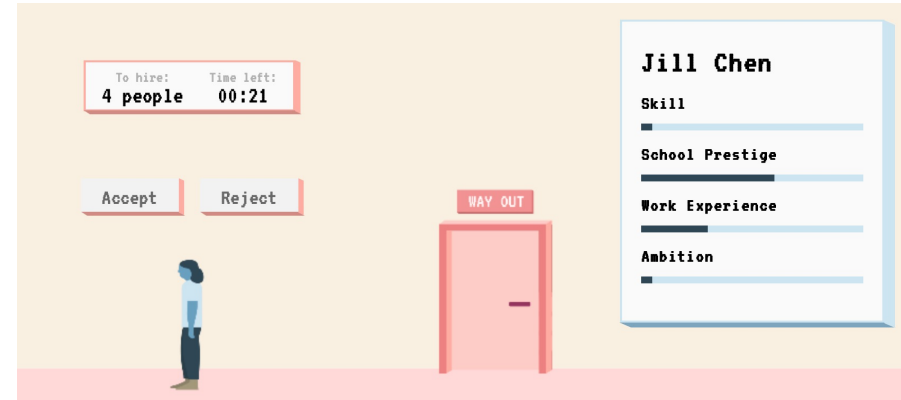
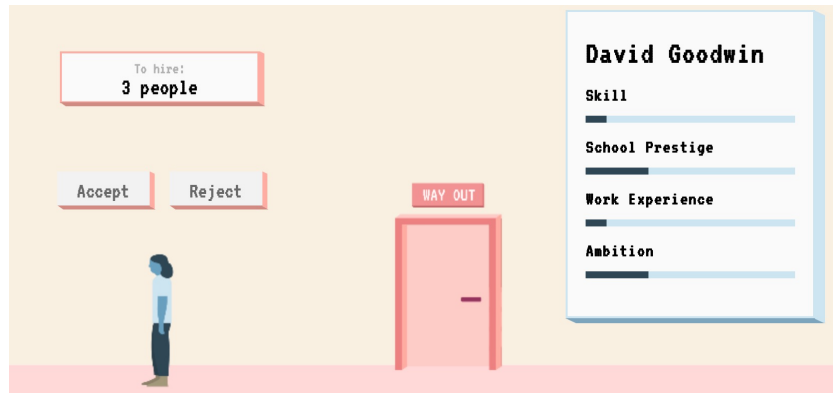
A **distribution** of those goods is fair if and only if it results from some (morally justifiable) process or procedure

(e.g., everyone *who is randomly chosen* to receive free healthcare for life, gets free healthcare for life)

# Applying those concepts to SOTBF

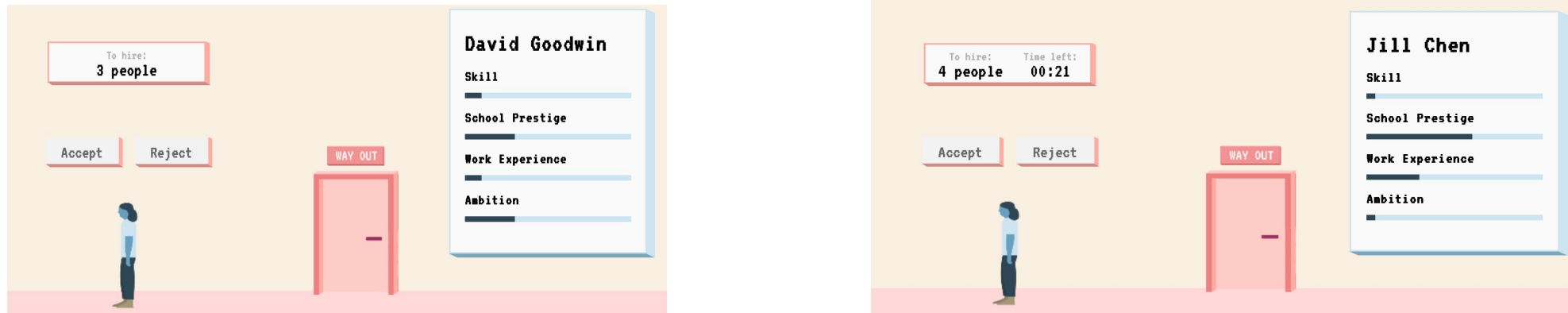


# Directly v indirectly disparate treatment (attributes v proxies)



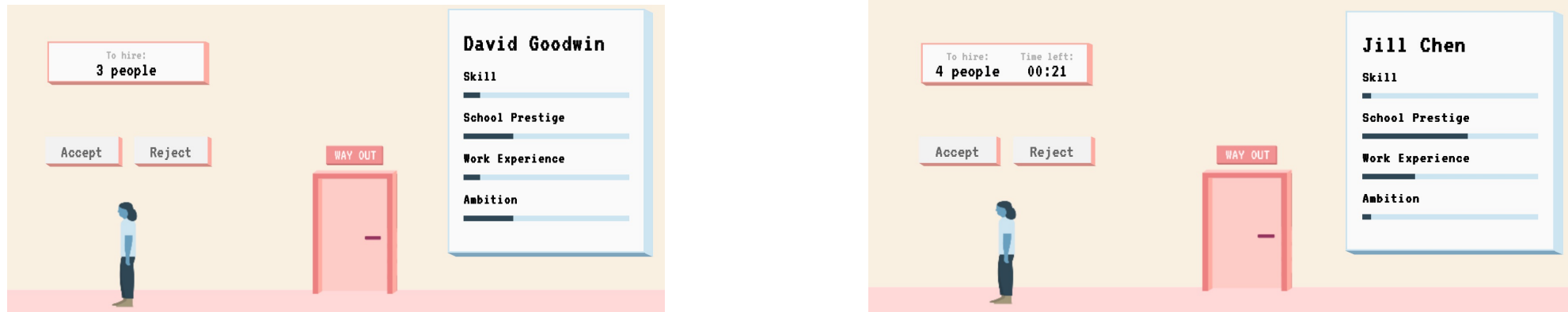


# Directly v indirectly disparate treatment (attributes v proxies)



In this case, “Skill” is a proxy for blue, so even if we stop *explicitly and directly* screening out a disproportionate number of the blue applicants, we might still *implicitly and indirectly* screen them out by training our model to focus on “Skill”.

# But why *shouldn't* we focus on skill?



Depending on how the modelers define it, “Skill” will be backwards-looking: it’ll be an indicator of someone’s prior experiences and successes. It won’t tell us whether they had any chance to *get* those experiences in the first place.

# Cases of unfairness in ML algorithms

Here are three ways that ML algorithms that automate decision making may treat people unfairly:

1) **In their purpose (goals):** the algorithm is designed to achieve a goal that is *itself* illegitimate, because that goal relies on false assumptions or reinforces attitudes or patterns of unjustified inequality

(continued) Here are three ways that ML algorithms that automate decision making may treat people unfairly:

2) **In their data collection practices (training data):** the algorithm is not as *accurate* as it could be because of poorly chosen target variables, underlying bias reproduced in training examples, unrepresentative samples, or coarse features

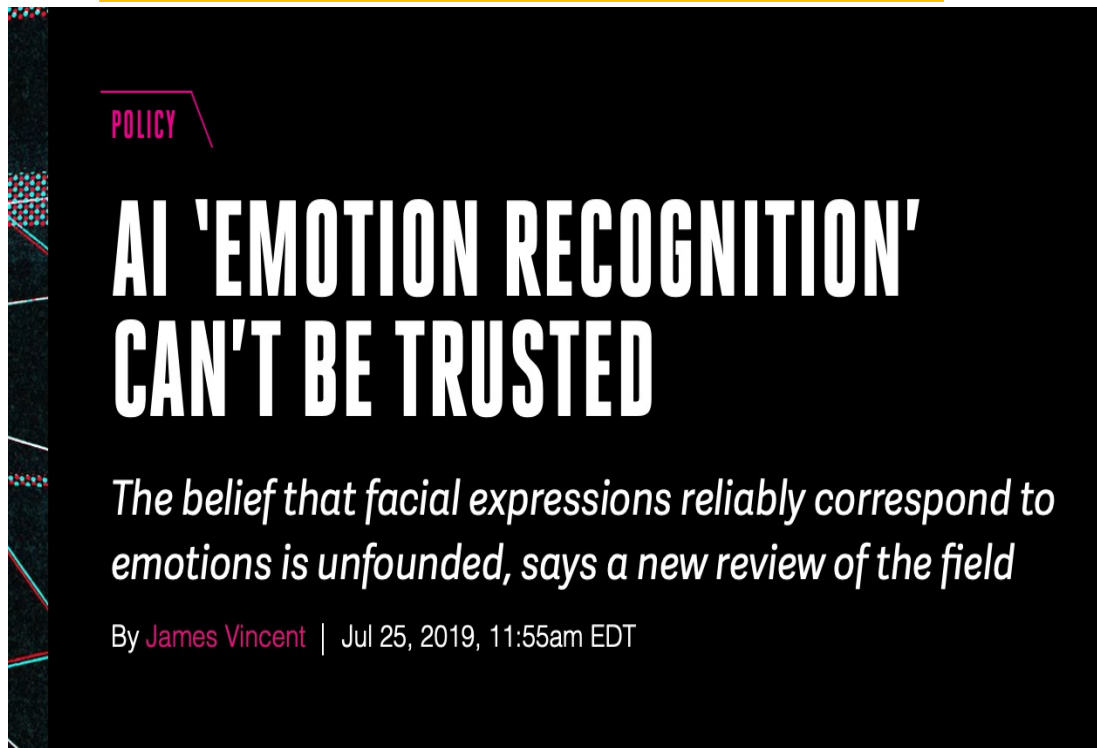
3) **In their distribution of burdens of error (outcomes):** the data and algorithm are as good as possible, but the algorithm imposes greater burdens of error on some stakeholders than others, often in ways that reinforce existing patterns of inequality in society

1. In Purposes (bad or flawed goal)

# Examples of Bad or Flawed Goals

- Ones based on empirically false assumptions
- Ones with a foreseeable high risk of making already-vulnerable groups even more vulnerable

# Example of Empirically False Assumptions



But the belief that we can easily infer how people feel based on how they look is controversial, and a significant new review of the research suggests there's no firm scientific justification for it.

"Companies can say whatever they want, but the data are clear," Lisa Feldman Barrett, a professor of psychology at Northeastern University and one of the review's five authors, tells *The Verge*. "They can detect a scowl, but that's not the same thing as detecting anger."



# Example of Increasing Vulnerability

---

From *The Guardian*,  
08 Sep 2017

The **research**, which went **viral** this week, used a sample of online dating photos, limited only to white users, to demonstrate that an algorithm could correctly distinguish between gay and straight men 81% of the time and 74% for women, suggesting machines can potentially have much better “gaydar” than humans.

The Human Rights Campaign (HRC) and GLAAD, two of the most prominent LGBTQ organizations in the US, slammed the study on Friday as “dangerous and flawed ... junk science” that could be used to out gay people across the globe and put them at risk. The advocates also criticized the study for excluding people of color and bisexual and transgender people and claimed the research made overly broad and inaccurate assumptions about gender and sexuality.

“It’s not biased”  $\neq$  “It’s morally harmless”

## 2. In Data Collection Practices (e.g., training data)

# Sources of bad or biased training data

- a. When defining target variables and in class labels
- b. When assembling the training data set, resulting in an unrepresentative sample
- c. When selecting relevant features
- d. Intentional bias: masking, redlining, etc.



## a. When defining target variables and class labels

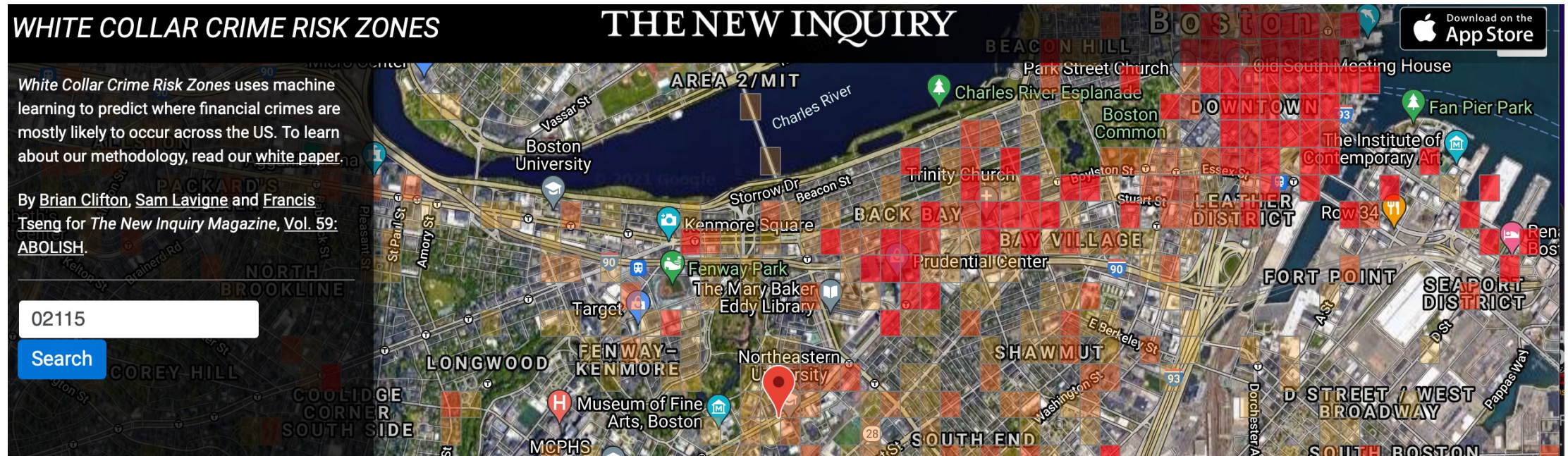
**Example: In the COMPAS case, the goal is to identify those at high risk of recidivism (reoffending). But what defines our target variable, “reoffending”?**

- Conviction of a new crime?
- Formal charges?
- Violation of probation?
- **Re-arrest (whether formally charged or not)? [this is what Northpointe chose]**





# How are the variables defined? (e.g., “crime”)



b. When assembling the training data set, resulting in an unrepresentative sample

- a) The **labeling** of training examples can be incautious or corrupted
- b) Data may be **inaccurate** or incomplete for certain classes of people, or it may **overrepresent** them

## a) Labeling of training examples can be incautious or corrupted

- **Example (mislabeling because of past prejudices):** St George's medical school applicants were rated by how similar they were to past students who were successful, but past students were selected partly as a result of sexist policies and practices (e.g., previously only males were admitted)
- **Example (mislabeling because of current prejudices):** LinkedIn's Talentmatch may learn whatever biases employers have, by adapting to clicks





b) Data may be inaccurate or incomplete for certain classes of people, or it may overrepresent (or underrepresent) them

- **Ex (institutions might maintain systematically keep less accurate, precise, timely or complete records for certain classes of people):** Credit reporting agencies
- **Ex (institutions might have records that overrepresent certain classes of people):** black and white people smoke marijuana at same rates (by own admission in surveys), but black people are 4-5 times more likely than white people to be arrested for marijuana-related offenses. In general, arrest rates are higher for black people than for white people in the U.S. (see next slides)

## c. When selecting relevant features

Ex in COMPAS case, which features of data should we select (ie, what data should we collect)?

(Publicly known) Criminal history?

Age at first arrest?

Gender?

Socio-economic status?

Current employment status?

“Criminal” companions/associations?

“Antisocial” behavior?

Family “criminality”?

DEPARTMENT OF JUSTICE BUREAU OF INVESTIGATION  
IDENTIFICATION CARD  
Division: S. Penitentiary, Atlanta, Ga. Located at: K4805

Received: MAY 4 1932  
From: A. J. Lee, Chicago  
Criminal: Vis. Income Tax Law  
Sentence: 10 yrs. 10 mos. 10 days  
Date of sentence: Oct 24-1921  
Sentence begins: May 4-1932  
Sentence expires: May 3-1942  
Good time sentence expires: Jan 19-1939  
Date of birth: 11-19 Occupation: Painter  
Birthplace: NY Nationality:  
Age: 35 Complexion: fair  
Height: 5-10 1/2 Eyes: grey  
Weight: 255 Hair: dark brown  
Build: stout

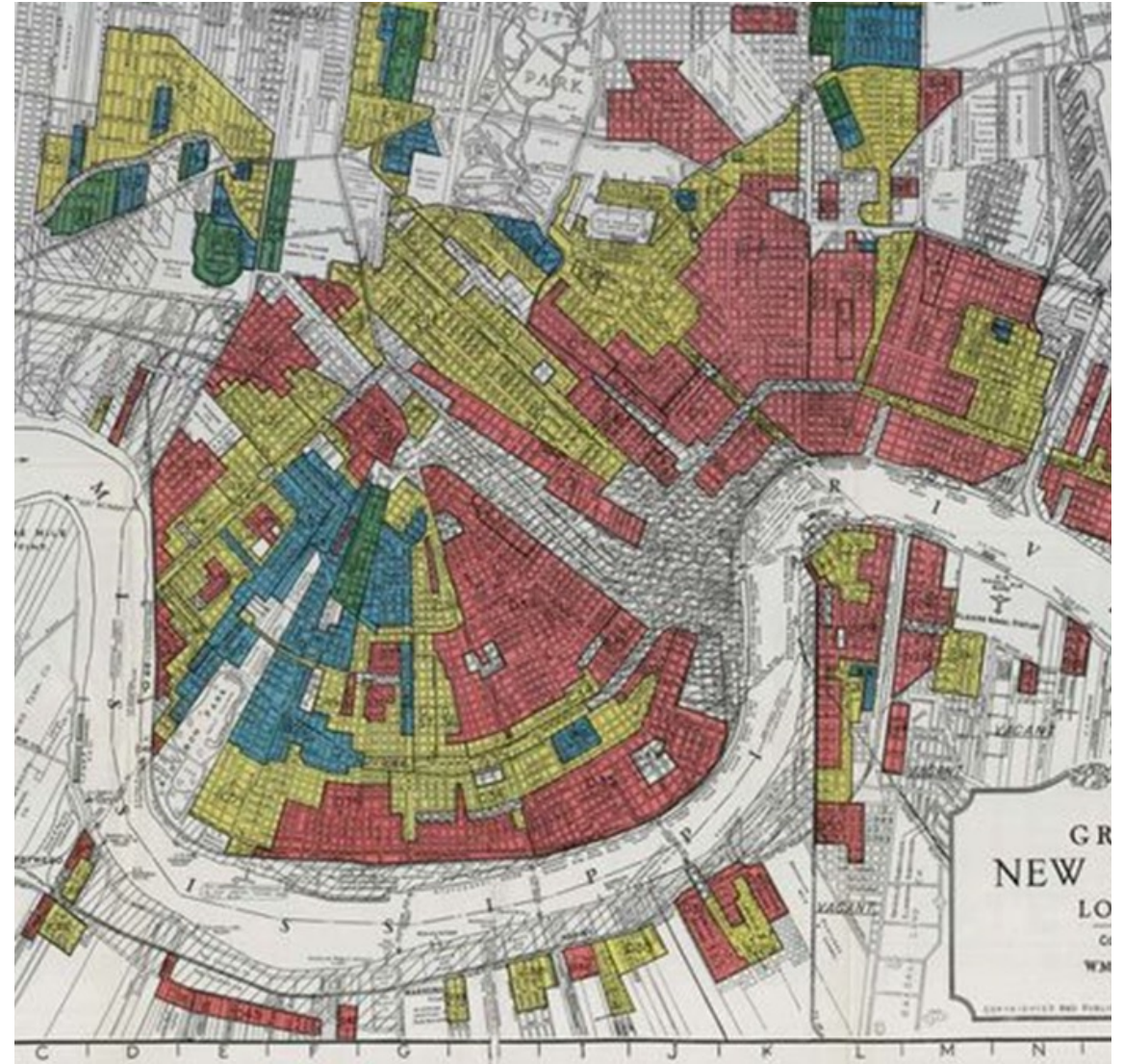
Scars and marks: oblique scar of 1/2" across cheek 2" in front left ear vertical scar of 2 1/2" on left jaw oblique scar of 2 1/2" under left ear on neck

| NAME | NUMBER      | CITY OR INSTITUTION | DATE        | CHARGE             | DISPOSITION OR SENTENCE |
|------|-------------|---------------------|-------------|--------------------|-------------------------|
| C    | NY City     | 1919                | Dis. Cond   | Discharged         |                         |
| D    | Chicago Ill | 1923                | Trapped     | Dismissed          |                         |
| E    | Do          | 5-8-24              | Murder 1st  | Released           |                         |
| F    | Do          | 6-7-26              | Vis. M.P.A. | Dismissed          |                         |
| H    | Do          | 7-28-26             | Murder      | Charged with death |                         |
| I    | Do          | 10-1-26             | Vis. M.P.A. | Dismissed          |                         |
| L    | Do          | 11-12-27            | Vis. M.P.A. | Dismissed          |                         |
| M    | Do          | 12-21-27            | Con. M.P.A. | Fined 26.00.00     |                         |
| N    | Do          | 5-17-29             | Do          | Dismissed          |                         |
| O    | Do          | 5-8-30              | Do          | Dismissed          |                         |

## d. *Intentional* bias in data collection: masking, redlining

### Data mining can obscure intentional discrimination

- Could intentionally bias data collection, refuse to audit it, or rely on coarse features (“digital redlining”)
- Data mining can enable institutions to circumvent legal barriers to unlawful discrimination by making proof very difficult to obtain



### 3. In Distribution of Burdens of Algorithmic Error (in decisions or outcomes)



# Case: COMPAS recidivism risk prediction algorithm



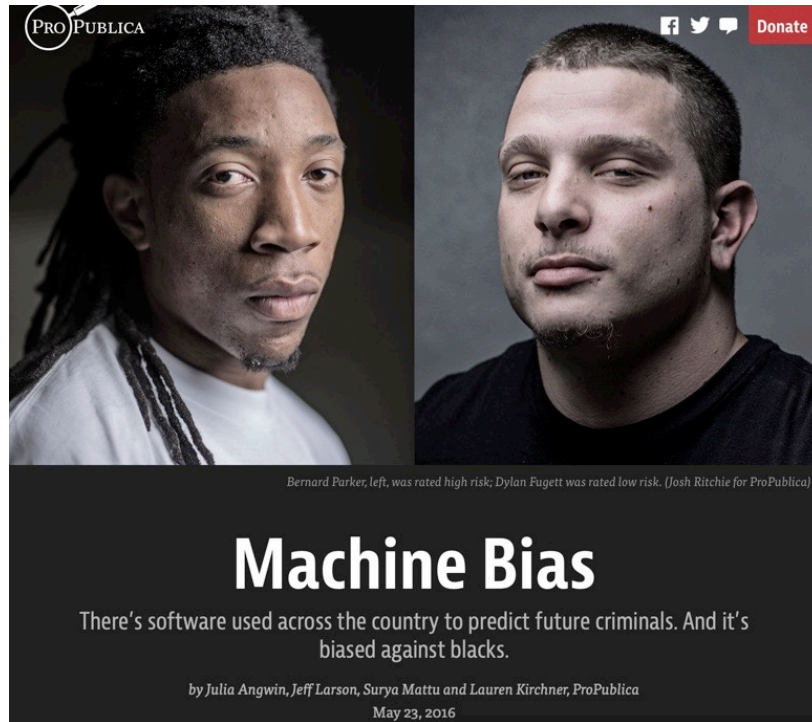
COMPAS generates a probability of re-arrest



Assigns a risk score of 1 to 10



Sorts into high (8-10), medium (5-7), or low (1-4) risk categories



|   | White | Black |
|---|-------|-------|
| Labeled Higher Risk & Didn't Re-Offend (False +s) | 23. % | 44.9% |
| Labeled Lower Risk & Did Re-Offend (False -s)     | 47.7% | 28.0% |

Unfair distribution of error by racial class membership

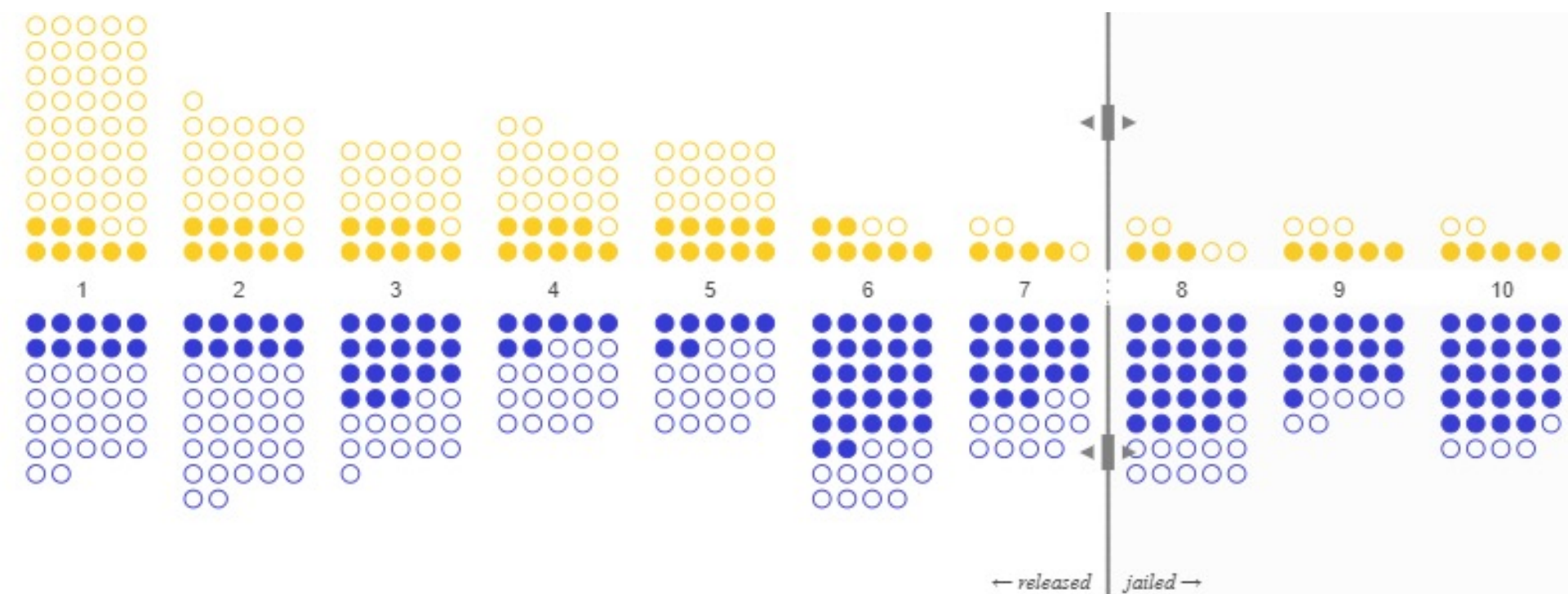
**ProPublica** examined COMPAS risk assessment scores for about 7000 defendants in Broward County, FL. Propublica found that while COMPAS correctly predicted recidivism 61% of the time, **the likelihood of different types of errors (FP and FNs) differed by race** (see table above)

**Northpointe** responded that COMPAS was nevertheless fair because its positive predictions of recidivism were correct at the same rates regardless of racial group membership; this response was followed by a detailed rebuttal by ProPublica

|                  |              | Actual Values |              |
|------------------|--------------|---------------|--------------|
|                  |              | Positive (1)  | Negative (0) |
| Predicted Values | Positive (1) | TP            | FP           |
|                  | Negative (0) | FN            | TN           |

Class Activity: [Can you make AI fairer than a judge?](#)

---



*black defendants*





# Two conceptions of accuracy?

$$P[Y = 1 | S > S_{HR}, r = w] = P[Y = 1 | S > S_{HR}, r = b]$$

*Predictive parity  
(Northpointe)*

$$P[S > S_{HR} | Y = 0, r = w] = P[S > S_{HR} | Y = 0, r = b]$$

*False positive parity*

$$P[S \leq S_{HR} | Y = 1, r = w] = P[S \leq S_{HR}, Y = 1 | r = b]$$

*False negative parity*

***Equalized odds  
(ProPublica)***

Some argue that these measurements each reflect different, incompatible concepts of fairness. What do you think?

***How should the risks of algorithmic error should be distributed?***

# ML and Treating People as Data Subjects

## The tension:

**“constructing the human as a data point for machine training and optimization rather than as a person who should be justly, equitably, and sensitively treated”**


(Chancellor et al., p 2)

# Group activity: KHASHM



Summing Up: some ways to  
address unfairness in ML  
algorithms

How do we avoid (creating or relying on machine learning algorithms that end up) treating people unfairly?



# First: explain whether an ML-based algorithm *should be used at all* in this domain

For all these reasons, there's a growing recognition among scholars and advocates that some biased AI systems should not be "fixed," but abandoned. As co-author Meredith Whittaker said, "We need to look beyond technical fixes for social problems. We need to ask: Who has power? Who is harmed? Who benefits? And ultimately, who gets to decide how these tools are built and which purposes they serve?"


**From Vox, "Some AI just shouldn't exist", 19 April 2019**

---



## Second: only if using an ML-based algorithm IS justified, then pay careful attention to how training data is collected

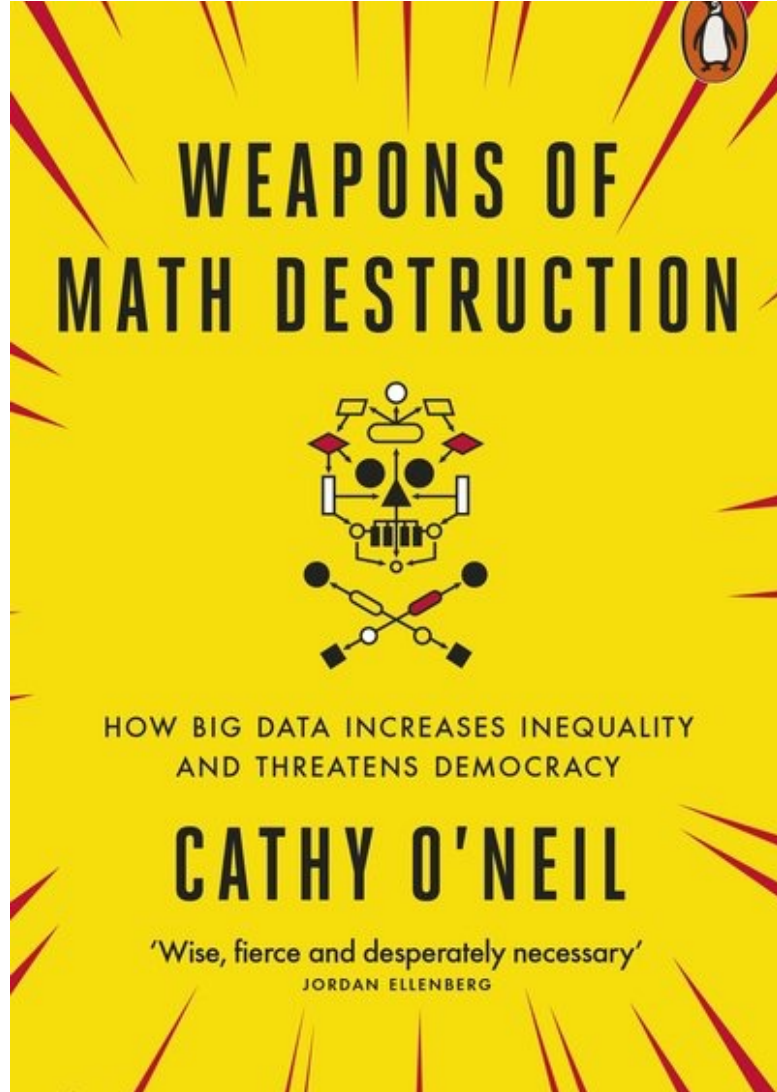
- When defining target variables and in class labels
  - When assembling the training data set, resulting in an unrepresentative sample
  - When selecting relevant features
  - Watch out for intentional bias: masking, redlining, etc.
-



Third: make  
*explicit* ethical  
decisions about  
how to distribute  
the risks/results  
of algorithmic  
error

- Even if the algorithm is perfectly accurate (?), there will be some unfairness in its results
- To distribute the risks of error ethically, you should bring in all stakeholders (designers, users, “bystanders”, etc.) for collaboration about definition/selection of relevant features, refinement of the training data set, etc.





We must, therefore, make careful, explicit choices as to how and where to distribute the burdens of error in the algorithms we build.

This should be done at both the **law and policy** level, and at the **design** level, which is where value-sensitive design – an approach that emphasizes stakeholder interests and values – attempts to intervene.

We should also ask *whether an algorithm should be used at all* for the task at hand.

# For Next Time

## Wednesday

- Quick review of Monday's materials
- Articulating values and identifying stakeholders: using value-sensitive design (VSD) to address unfairness in ML
- Revisiting SOTBF and KHASM
- Conclusion: Keeping the human in machine learning

# Thank you!

## Some review questions:

- What does it mean to treat people fairly?
  - What are the three main ways that algorithms that automate decision-making might treat people unfairly?
  - Describe some of the ways that training data might be corrupted or biased and so result in unfair treatment.
  - What is predictive parity, as opposed to error rate balance (equalized odds)?
  - Why are there necessarily trade-offs between these measures of fairness in algorithmic design?
  - How should we deal with such trade-offs? What should we do about them?
-