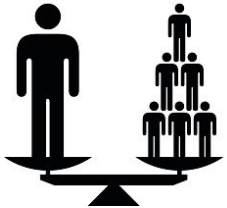
Fairness - Tradeoffs

Group 3

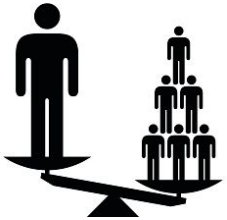
Zhenyu Lei, Leena Bacha, Brooke Hewitt,
Mihika Rao, Jett Yan

What Does It Mean for an Algorithm to Be Fair?

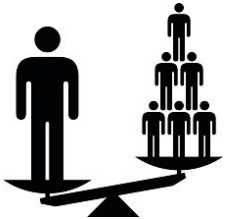
- To what extent should factors outside of an individual's control be factored into decisions made about them?



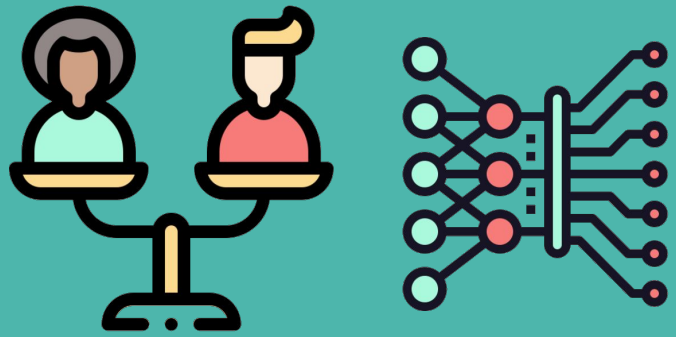
- Equality of outcomes vs. equality of treatment



- Can fairness be guaranteed?



Approaches Toward Defining Algorithmic Fairness



Three Key Spaces in Algorithmic Decision-Making

- ***Construct Space:*** The space containing the features we would like to make a decision based on (ex. intelligence)
- ***Observed Space:*** Measurable features that can be observed (ex. SAT score)
- ***Decision Space:*** The final decision output(s) (ex. College admission decision)
- Ex. College Admissions Process

Decision space	Construct space	Observed space
Performance in college	Intelligence	IQ
Performance in college	Success in High School	GPA
Recividism	Propensity to commit crime	Family history of crime
Recidivism	Risk-averseness	Age
Employee Productivity	Knowledge of job	Number of Years of Experience

S.A. Friedler, C. Scheidegger, S. Venkatasubramanian, 2017

Gromov-Wasserstein Distance (GWD)

- Computes the distance between the sets of pairs of points to determine whether the two point sets determine similar sets of distances

- Used to compare data from groups that may have different metric spaces
  - Ex. Medical images from different devices
  - Ex. Comparing hiring data across groups with different evaluation metrics

$$\mathcal{GW}(X, Y) = \frac{1}{2} \inf_{\nu \in \mathcal{U}(X,Y)} \int \int |d_X(x, x') - d_Y(y, y')| d\mu_X \times d\mu_X d\mu_Y \times d\mu_Y$$

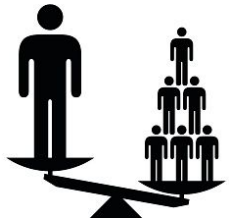
Sets of Points

Respective Distance Functions

Probability Measures

S.A. Friedler, C. Scheidegger, S. Venkatasubramanian, 2017

What You See Is What You Get (WYSIWYG) Worldview

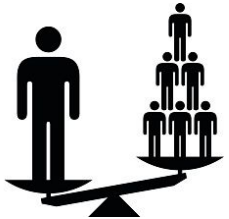


- A decision process is fair if ***individuals*** close in the construct space receive similar outcomes in the decision space

- Asserts that the construct and observed spaces are essentially the same

S.A. Friedler, C. Scheidegger, S. Venkatasubramanian, 2017

Structural Bias Worldview



- A decision process is fair if ***groups*** receive statistically similar outcomes to avoid discrimination
  - “We’re all equal” (WAE) mindset

- ***Group:*** A collection of individuals that share a certain set of characteristics

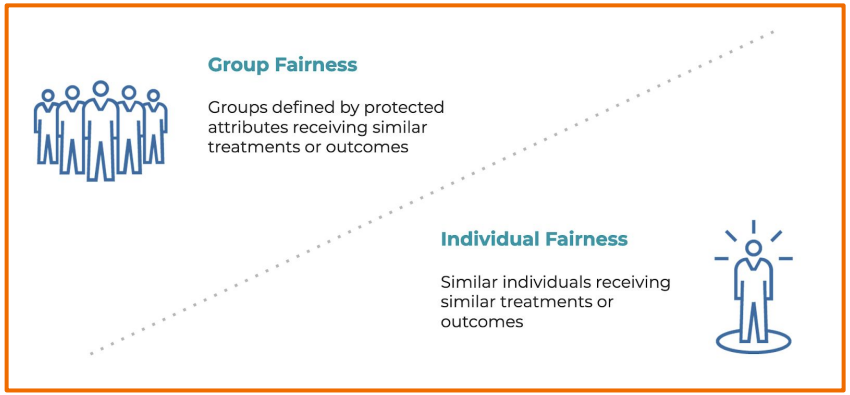
- ***Structural Bias:*** More distortion between groups than there is within groups when mapping between the construct and observed space

S.A. Friedler, C. Scheidegger, S. Venkatasubramanian, 2017

Individual Fairness vs. Group Fairness

- ***Under WYSIWYG worldview:*** fairness can be guaranteed
  - Group fairness mechanism would be unfair in this worldview

- ***Under structural bias worldview:*** non-discrimination can be guaranteed
  - Applying an individual fairness mechanism will cause discrimination in the decision space in this worldview



S.A. Friedler, C. Scheidegger, S. Venkatasubramanian, 2017

Mathematical Definitions of Algorithmic Fairness

Statistical Parity

When an equal proportion of defendants are detained in each race group

$$\mathbb{E}[d(X) \mid g(X)] = \mathbb{E}[d(X)]$$

Used when the focus is on **overall representation and equal outcomes** across groups. Ensures that the outcomes of the algorithm are balanced across different groups, without considering other factors

Limitation: It **doesn't account for individual risk or other relevant factors**, potentially leading to unfair outcomes for individuals

Conditional Statistical Parity

Controlling for a limited set of “legitimate” risk factors, an equal proportion of defendants are detained within each race group

$$\mathbb{E}[d(X) \mid \ell(X), g(X)] = \mathbb{E}[d(X) \mid \ell(X)]$$

Used when there are **legitimate factors that should be considered** in the decision-making process, and **fairness** should be **ensured within subgroups** defined by these factors

Limitation: It relies on the **identification and selection of “legitimate” factors,** which can be subjective may still mask underlying biases. It is difficult to restrict to legitimate features that do not correlate with race

S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, 2017

Predictive Equality

The accuracy of decisions is equal across race groups, as measured by the false positive rate (FPR)

$$\mathbb{E}[d(X) \mid Y = 0, g(X)] = \mathbb{E}[d(X) \mid Y = 0]$$

When the goal is to **minimize false positives** and ensure that the **algorithm is equally accurate across groups** in identifying individuals who should not be subject to the adverse outcome

Limitation: Does not address **disparities in false negative rates** and may lead to unequal outcomes in terms of overall detention rates

S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, 2017

Immediate Utility

Decision metric that **captures** the proximal **costs and benefits** of a decision rule

$$
\begin{aligned}
u(d, c) &= \mathbb{E}\left[\mathbb{E}\left[Yd(X) - cd(X) \mid X\right]\right] \\
&= \mathbb{E}\left[p_{Y|X} d(X) - cd(X)\right] \\
&= \mathbb{E}\left[d(X)(p_{Y|X} - c)\right]
\end{aligned}
$$

Used to balance the prevention of violent crime committed by released defendants against the costs of detention

Limitation: All violent crime is assumed to be **equally costly**. The cost of detaining every individual is assumed to be *c*, **without regard to personal characteristics**

S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, 2017

Mathematical Outline for Algorithmic Fairness

- ***Statistical Parity**:* Equal proportion of defendants are detained in each race group
- ***Conditional Statistical Parity**:* Controlling for a limited set of "legitimate" risk factors, an equal proportion of defendants are detained within each race group
- ***Predictive Equality**:* The accuracy of definitions is equal across race groups, as measured by the false positive rate (FPR)
- ***Immediate Utility**:* Decision metric that captures the proximal costs and benefits of a decision rule.


S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, 2017

- ***Statistical Parity***:

$$\mathbb{E}[d(X) \mid g(X)] = \mathbb{E}[d(X)]$$

- ***Conditional Statistical Parity***:

$$\mathbb{E}[d(X) \mid \ell(X), g(X)] = \mathbb{E}[d(X) \mid \ell(X)]$$

- ***Predictive Equality***:

$$\mathbb{E}[d(X) \mid Y = 0, g(X)] = \mathbb{E}[d(X) \mid Y = 0]$$

- ***Immediate Utility***:

$$\begin{aligned} u(d,c) &= \mathbb{E}\left[\mathbb{E}\left[Yd(X) - cd(X) \mid X\right]\right] \\ &= \mathbb{E}\left[p_{Y|X}d(X) - cd(X)\right] \\ &= \mathbb{E}\left[d(X)(p_{Y|X} - c)\right] \end{aligned}$$

S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, 2017

Assumptions for an Optimal Decision

- Policymakers would prefer to ***maximize immediate utility*** as the benefits outweigh the costs

- All violent crimes are assumed to be equally costly because the benefit of detaining a defendant who would have committed a violent crime if released is binary

- The cost of detaining every individual is assumed to be c, regardless of personal characteristics

S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, 2017

Optimal Decision Rule

Goal: Maximize immediate utility subject to fairness constraints

- ***Unconstrained Optimum****:* The algorithm deterministically detains defendants if and only if $pY|X \geq c$. In this case, a single, uniform threshold is applied to all individuals, irrespective of group membership

- ***Statistical Parity:*** Detain individuals if and only if $pY|X \geq tg(X)$, where $tg(X)$ is a threshold that depends only on group membership

- ***Predictive Equality****:* Similar to statistical parity, the optimal rule detains defendants based on group-specific threshold

- ***Conditional Statistical Parity****:* The optimum is to detain individuals if and only if $pY|X \geq tg(X)|\ell(X)$ where $tg(X)\ell(X)$ represents group membership and legitimate attributes

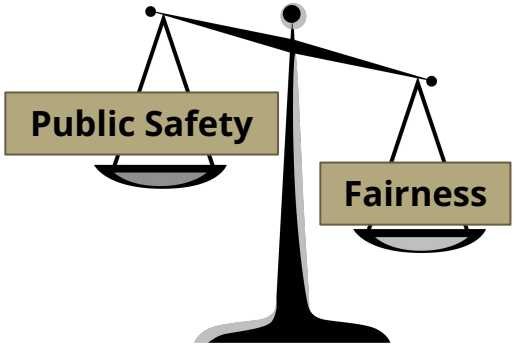
S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, 2017

The Impossibility Result

- Tension between satisfying common fairness constraints and treating all individuals equally, irrespective of race, since the optimal constrained algorithms differ from the optimal unconstrained algorithm

- No algorithm can simultaneously satisfy calibration, balance for negative class, and balance for positive class

- Removing fairness constraints allows for a single optimal decision threshold, maximizing public safety but still leading to racial disparities

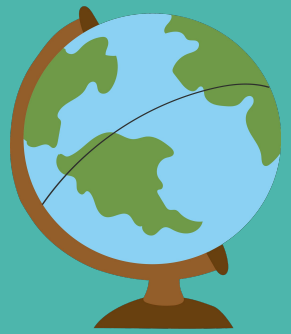


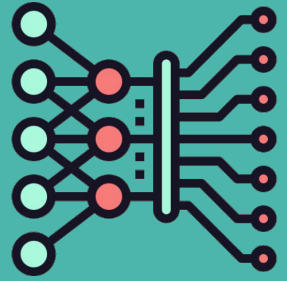
S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, 2017

Algorithmic Fairness - Critical Findings and Analysis

- There are different types of fairness and some notations of fairness are incompatible with each other
- Fairness cannot be universally defined as different worldviews conflict, yet researchers must be explicit about their assumptions when designing these systems
- Assumptions about fairness worldviews may reflect biases in society
- Assumes that fairness can be mathematically formalized, but in practice, construct spaces are difficult to define and measure

Real-World Applications of Algorithmic Fairness





Fairness in Real-World Problems

What does fairness mean rigorously?

- ***COMPAS***: Used in courts to predict recidivism, criticized for racial bias.
- ***Healthcare***: Diagnostic tools perform differently across genders/races.
- ***Ads***: Job ads shown disproportionately to certain demographics.

J. Kleinberg, S. Mullainathan, M. Raghavan, 2017

Three Fairness Criteria

1. ***Calibration*** $\text{Score } v_b \text{ in group } t \implies \mathbb{P}(\text{ Outcome }) = v_b$

$$\mathbb{E}\left[Y \mid v_b,\ \text{Group } t\right] = v_b$$

***Example*:** A risk score of 0.7 means 70% of people with that score reoffend—**for every group**.

2. ***Balance for Positive Class***

Average score for people who do reoffend is equal across groups.

$$\frac{\sum \text{Scores for positives in Group 1}}{\mu_1} = \frac{\sum \text{Scores for positives in Group 2}}{\mu_2}$$

3. ***Balance for Negative Class***

Average score for people who don't reoffend is equal across groups.

$$\frac{\sum \text{Scores for negatives in Group 1}}{N_1 - \mu_1} = \frac{\sum \text{Scores for negatives in Group 2}}{N_2 - \mu_2}$$

J. Kleinberg, S. Mullainathan, M. Raghavan, 2017

The Impossibility Theorem

***No risk score can satisfy all three fairness criteria unless*:**

Perfect Prediction $(p_\sigma \in \{0, 1\})$

- All predictions are deterministic

Equal Base Rates $(\mu_1/N_1 = \mu_2/N_2)$

- Groups have identical prevalence of the outcome.

$$\frac{\mu_1(1-\gamma)}{N_1-\mu_1} = \frac{\mu_2(1-\gamma)}{N_2-\mu_2} \implies \gamma = 1 \text{ or } \frac{\mu_1}{N_1} = \frac{\mu_2}{N_2}$$

J. Kleinberg, S. Mullainathan, M. Raghavan, 2017

Approximate Fairness?

Even allowing small errors, systems must resemble one of the two edge cases.

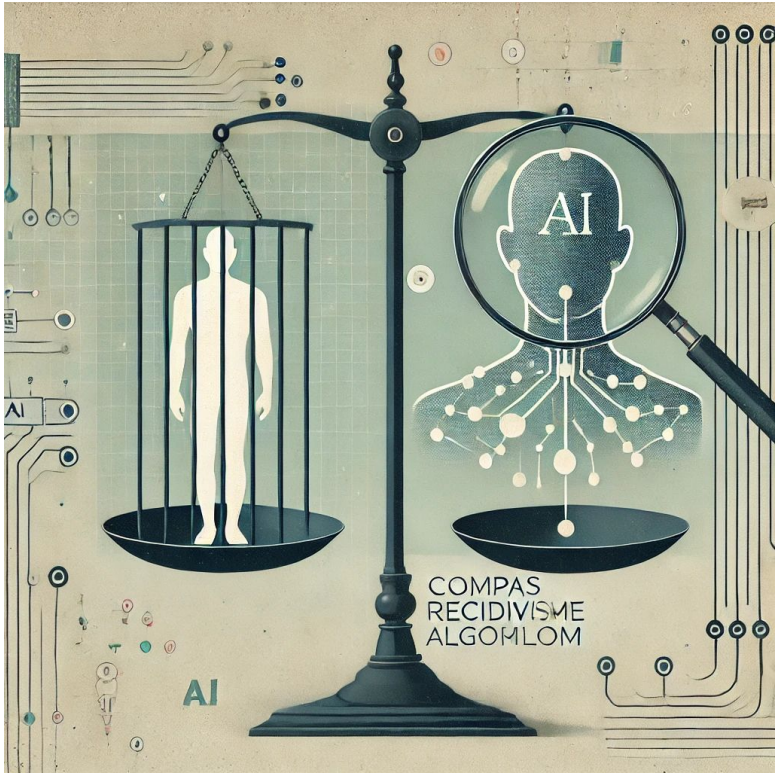
$$\text{Unfairness} \propto \left| \frac{\mu_1}{N_1} - \frac{\mu_2}{N_2} \right|$$

- ***Small relaxations don't eliminate trade-offs—they just make them less severe.***

J. Kleinberg, S. Mullainathan, M. Raghavan, 2017

COMPAS

- ***Risk assessment tool*** that predicts a defendant's recidivism (likelihood of reoffending)
- Decision based on over 100 factors including age, sex, and criminal history
- Found to be less accurate than random untrained human evaluators
- Revealed that black defendants are substantially more likely to be classified as high risk



S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, 2017

Bias in COMPAS

Among defendants who ultimately did not re-offend, ***black defendants were more than twice as likely as white defendants to be labeled as risky***

	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%

Source ProPublica analysis of data from Broward County, FL

Analyze Fairness in RPIs

- ***Risk Prediction Instrument (RPI)***

- ***Test Fairness Definition***:

  A risk score is considered fair if it predicts the likelihood of recidivism equally across different groups

$$\mathbb{P}(Y = 1 \mid S = s, R = b) = \mathbb{P}(Y = 1 \mid S = s, R = w)$$

- ***Key Error Metrics***:
  - False Positive Rate (FPR): The probability of incorrectly classifying a non-recidivist as high-risk
  - False Negative Rate (FNR): The probability of incorrectly classifying a recidivist as low-risk
  - Positive Predictive Value (PPV): The probability that an individual predicted as high-risk actually reoffends
  - Prevalence (p): The base rate of recidivism in the population

- ***Relationship Between FPR, FNR, PPV, and p:***

$$FPR = \frac{p}{1-p} \cdot \frac{1 - PPV}{PPV} \cdot (1 - FNR)$$

When a model is fair in predictive probabilities, differences in recidivism rates (p) across demographic groups will lead to unequal FPR and FNR, which can result in disparate impacts

A. Chouldechova, 2016

Impact Assessment

$$T_{\text{MinMax}} = \begin{cases} t_L & \text{if } S_c = \text{Low-risk} \\ t_H & \text{if } S_c = \text{High-risk} \end{cases}$$

- ***For Non-Recidivators:***

$$\Delta = (t_H - t_L) \cdot (FPR_b - FPR_w)$$

Higher false positive rates (FPR) for black defendants lead to harsher penalties for innocent individuals, even though they pose no real risk.

- ***For Recidivators:***

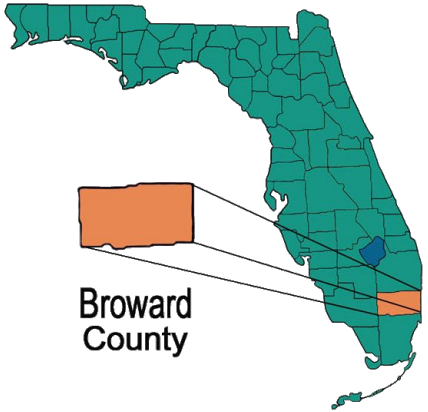
$$\Delta = (t_H - t_L) \cdot (FNR_w - FNR_b)$$

Higher false negative rates (FNR) for white defendants result in lenient sentencing for individuals who genuinely pose a risk, while Black defendants face stricter penalties.

A. Chouldechova, 2016

COMPAS Comparison

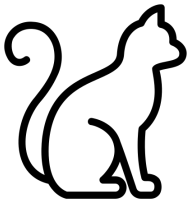
- Retrained a model using Broward County Data to investigate the practical implications of the trade-off between fairness and public safety
- In retrained model, estimated: (1) the increase in violent crime from releasing more high-risk defendants and (2) the proportion of detained defendants who are actually low-risk
- Risk assessment model outperformed COMPAS (0.75 vs. 0.73), but enforcing fairness constraints leads to unintended consequences
- Ensuring fairness results in an increase in violent recidivism while also detaining more low-risk defendants



S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, 2017

Real-World Example

- ***ProPublica Analysis:*** COMPAS is calibrated but fails balance.



- **Black defendants:**
  Higher false positives.
- **White defendants:**
  Higher false negatives.

COMPAS is mathematically fair in one sense (calibration) but unfair in another (error rates).

J. Kleinberg, S. Mullainathan, M. Raghavan, 2017

Ethical Design Choices

Transparency: Disclose which fairness criteria are relaxed.

Context Matters:

- Criminal justice: Prioritize calibration (truthful risks).
- Ads: Prioritize balance (avoid stereotyping).

Regulation: Mandate audits for chosen fairness criteria.

J. Kleinberg, S. Mullainathan, M. Raghavan, 2017

Paper Findings/Critiques

- These papers provide a **foundation** for comparing and contrasting these different notions of fairness by developing metrics toward defining algorithmic fairness

- No paper was able to come with an optimal solution for fairness without sacrificing public wellbeing

- ***Lack of Factor Analysis:*** Does not identify which factors were most accurate of indicative of risk scores, instead placing emphasis on race

- ***Limited Scope:*** Lack of case studies

Q&A

1. How can modifying an algorithm to enhance fairness impact its outcomes? What are the potential benefits and trade-offs?

2. Can fairness be redefined to avoid trade-offs?

3. Is it possible to avoid trade-offs by using more data or different features?

4. Should AI algorithms be applied to help/support legal decisions?

5. Since AI algorithms are trained on data - to what extent should factors outside of an individual's control be factored into decisions made about them?
    a. Ex. COMPAS recidivism rating

Case Study: Bias in Resume Screening Algorithms

Scenario: You are part of a recruitment team at a tech company that uses an AI tool to help screen resumes for job applications. After an audit, you discover the following biases:

- White-associated names were selected 85% of the time for interviews, while black-associated names were selected only 9% of the time.

- Male associated names were chosen 52% of the time, even in female-dominated fields.

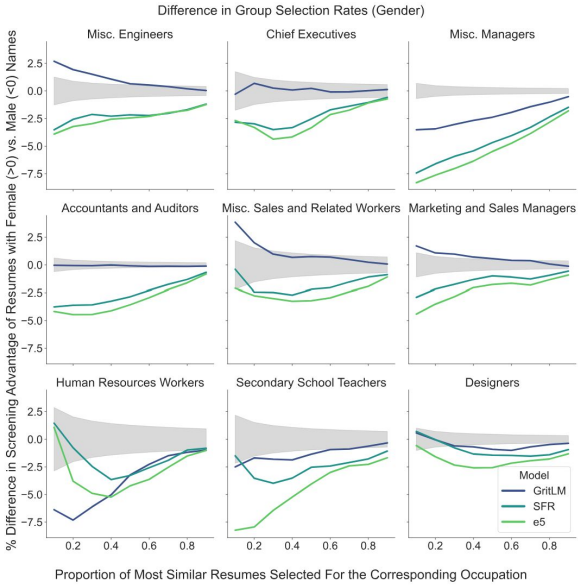
- Black men had their resumes overlooked 100% of the time compared to other candidates.

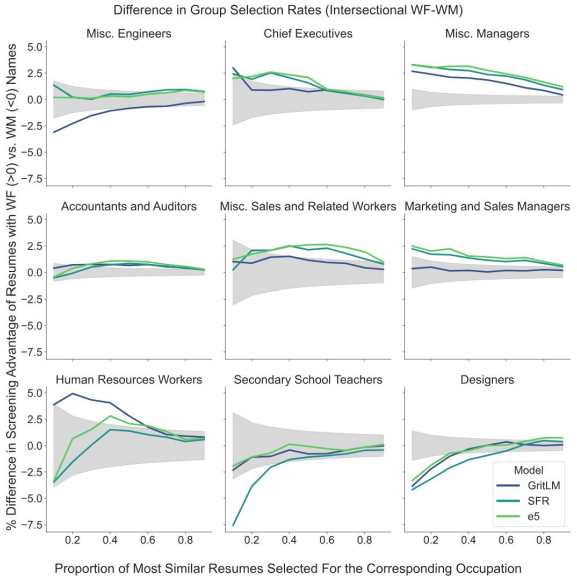
Case Study: Bias in Resume Screening Algorithms

Discussion questions:

- What actions would you take to address this bias?
- Would increasing the data set improve fairness?
- Should hiring decisions combine AI recommendations with human oversight?
- How can you ensure the training data is balanced?

Case Study: Bias in Resume Screening Algorithms





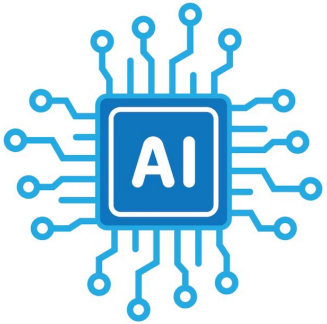
Case Study Audit? Possibly

https://ojs.aaai.org/index.php/AIES/article/view/31748/33915

Resume analyzer model that has bias: some main points:

- Resumes with White-associated names were selected 85% of the time for the next hiring step, while resumes with Black-associated names were only preferred 9% of the time.
- Resumes with male-associated names were preferred 52% of the time, even for roles with a traditionally high representation of women – like HR positions (77% women) and secondary school teachers (57% women).
- Resumes with White female names were chosen over those with Black female names, by a margin of 48% to 26%.
- Black men faced the greatest disadvantage, with their resumes being overlooked 100% of the time in favor of other candidates.
- We could show images of this results and see if people can see the bias and what can be done?

What Can Be Done to Avoid Discrimination?



- Audit AI models for bias on a regular basis (can be costly)
- Ensure that data used to train models is balanced and representative, and prioritizes models with built-in transparency (feeding in ***data influenced by historic inequalities*** can result in decisions that are inherently flawed or biased)
- Integrate and mandate human oversight into AI decisions
- Ensure selection criteria is neutral and inclusive