Fairness In AI

Alex Wollam and David Sarpong

Introduction: Fairness In Society

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

Discussion #1

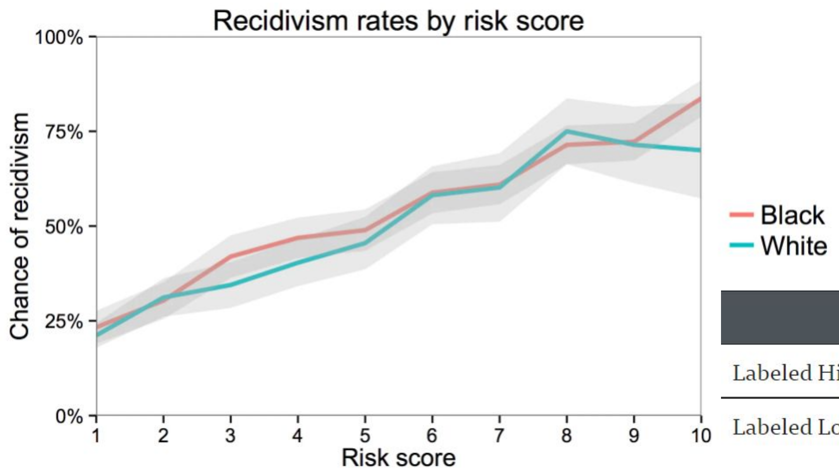
Defining Fairness

Q1: Briefly in your own words, describe what it means to you for something to be fair. What properties might this have?

Q2: The previous anecdotes don’t seem very fair. How does it violate your definition?

Q3: Can you think of other ways in which fairness could be violated?

COMPAS Fairness



	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: A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

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

What is Fairness

- ProPublica authors argue imbalanced risk scores (classification) in each group

- Rebuttal: Scores are well-calibrated; i.e., if there is a 60% of recidivism; 60% of observed persons re-offend.

Which definition of fairness do we use?

Proposed Metrics for Fairness

Some measures of fairness are:

- Calibration within groups
- Balance for the positive class
- Balance for the negative class

Some measures of unfairness are:

- Disparate treatment
- Disparate impact
- Disparate mistreatment

Inherent trade-offs in the fair determination of risk scores. Kleinberg et al.

Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. Zafar et al.

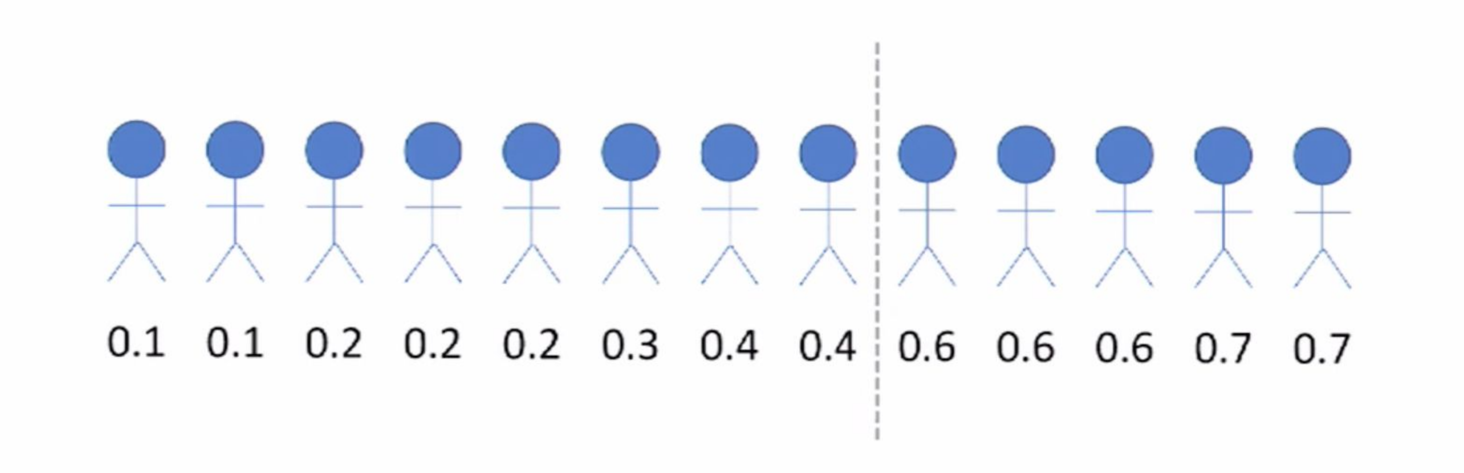
Proposed Metrics for Fairness

- Calibration within groups:
    - If the algorithm identifies a set of people as having a probability z of constituting positive instances, then approximately a z fraction of this set should indeed be positive instances
- Balance for the positive class:
    - The average score received by people constituting positive instances should be the same in each group
- Balance for the negative class
    - The average score received by people constituting negative instances should be the same in each group

Inherent trade-offs in the fair determination of risk scores. Kleinberg et al.

Calibration Within Groups

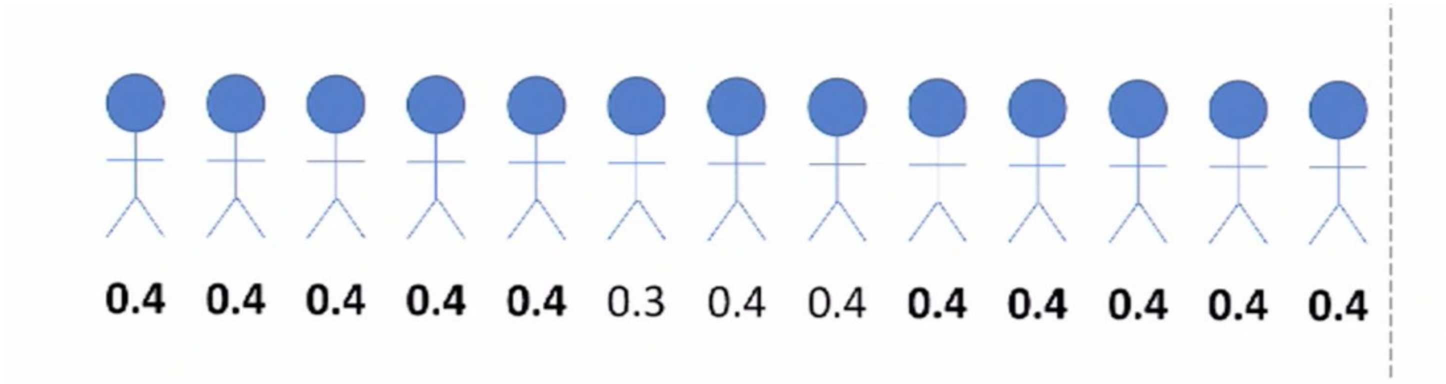
Before calibration:



Source of Image: NIPS'17 Tutorial on [Fairness in Machine Learning].

Calibration Within Groups

After calibration:

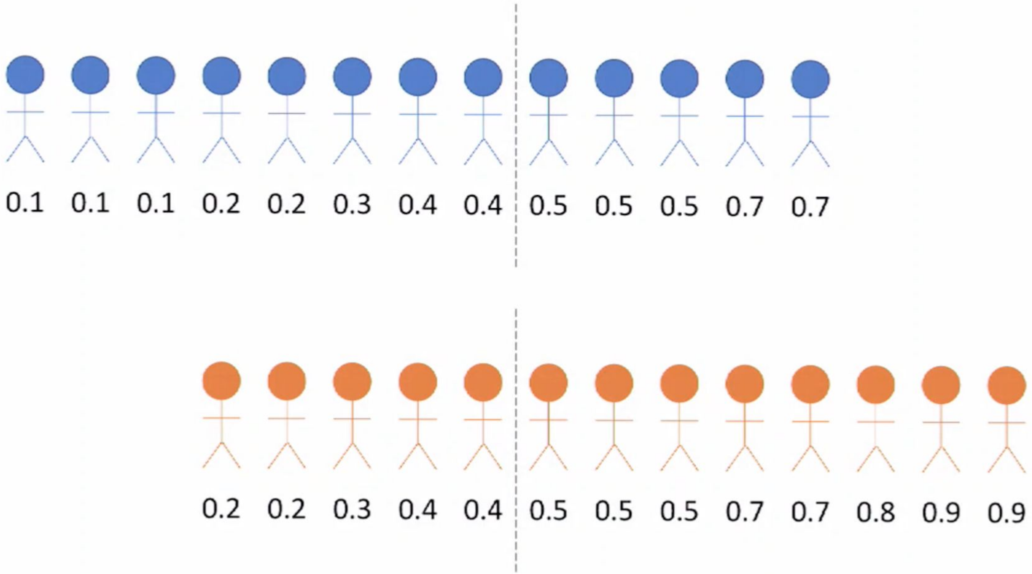


Source of Image: NIPS’17 Tutorial on Fairness in Machine Learning.

Balance For Positive (& Negative) Class

Two (sensitive) groups: blue & orange

Imbalance between groups



Source of Image: NIPS’17 Tutorial on Fairness in Machine Learning.

Proposed Metrics for Unfairness

- Disparate treatment:
  - The probability in predicting a specific label y given a feature x changes after observing the sensitive feature z
  - $P(\hat{y}|\mathbf{x}, z) \neq P(\hat{y}|\mathbf{x})$
- Disparate impact:
  - The probability in assigning a user to the positive class, y = 1, is not the same across sensitive features z
  - $P(\hat{y} = 1|z = 0) \neq P(\hat{y} = 1|z = 1)$
- Lack of disparate mistreatment:
  - The misclassification rates for different groups of people having different values of the sensitive feature z are not the same

Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. Zafar et al.

Disparate Treatment: $P(\hat{y}|\mathbf{x}, z) \neq P(\hat{y}|\mathbf{x}),$

User Attributes		
Sensitive	**Non-sensitive**	
Gender	**Clothing Bulge**	**Prox. Crime**
Male 1	1	1
Male 2	1	0
Male 3	0	1
Female 1	1	1
Female 2	1	0
Female 3	0	0

Ground Truth (Has Weapon)
✓
✓
✗
✓
✗
✓

Classifier's Decision to Stop		
C_1	**C_2**	**C_3**
1	1	1
1	1	0
1	0	1
1	0	1
1	1	1
0	1	0

Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. Zafar et al.

Disparate Impact: $P(\hat{y}=1|z=0) \neq P(\hat{y}=1|z=1)$

User Attributes			Ground Truth (Has Weapon)	Classifier's Decision to Stop		
Sensitive	Non-sensitive					
Gender	Clothing Bulge	Prox. Crime		C_1	C_2	C_3
Male 1	1	1	✓	1	1	1
Male 2	1	0	✓	1	1	0
Male 3	0	1	✗	1	0	1
Female 1	1	1	✓	1	0	1
Female 2	1	0	✗	1	1	1
Female 3	0	0	✓	0	1	0

Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. Zafar et al.

Disparate Mistreatment

User Attributes		
Sensitive	Non-sensitive	
Gender	Clothing Bulge	Prox. Crime
Male 1	1	1
Male 2	1	0
Male 3	0	1
Female 1	1	1
Female 2	1	0
Female 3	0	0

Ground Truth (Has Weapon)
✓
✓
✗
✓
✗
✓

Classifier's Decision to Stop		
C_1	C_2	C_3
1	1	1
1	1	0
1	0	1
1	0	1
1	1	1
0	1	0

Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. Zafar et al.

Discussion #2

Fairness Tradeoffs

Q1: Given these different measures of fairness, what considerations should be made when choosing how to balance them?

Q2: What fairness measures do you think are most important for COMPAS?

Q3: Under these new considerations, do you now believe COMPAS to be fair or unfair?

Can We Develop Theory For Fairness In Data-Driven Systems?

A simple model to investigate fairness

Inherent trade-offs in the fair determination of risk scores. Kleinberg et al.

Tradeoffs in Fairness

Characterization Theorem:

It's impossible to satisfy fairness in all three "notions" of fairness non-trivially

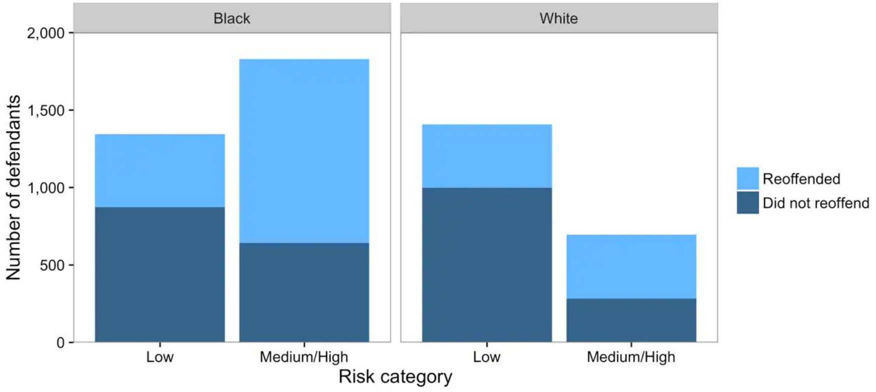
Pick at most two; drop at least one:

- Calibration within groups
- Balance for the positive class
- Balance for the negative class

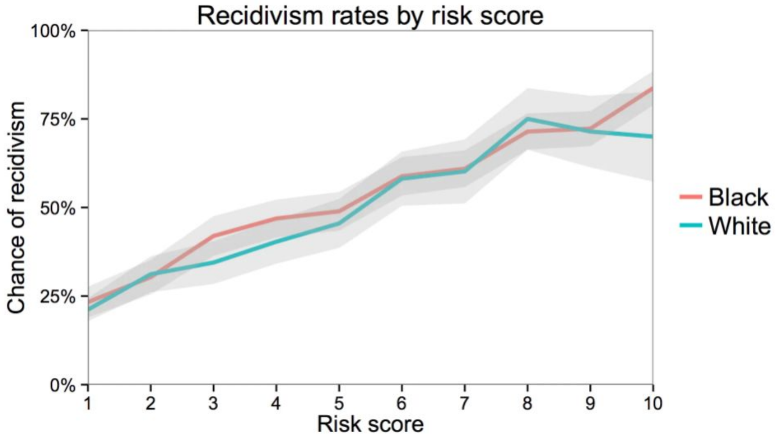
Inherent trade-offs in the fair determination of risk scores. Kleinberg et al.

Tradeoffs in Fairness

ProPublica complaint and the characterization theorem



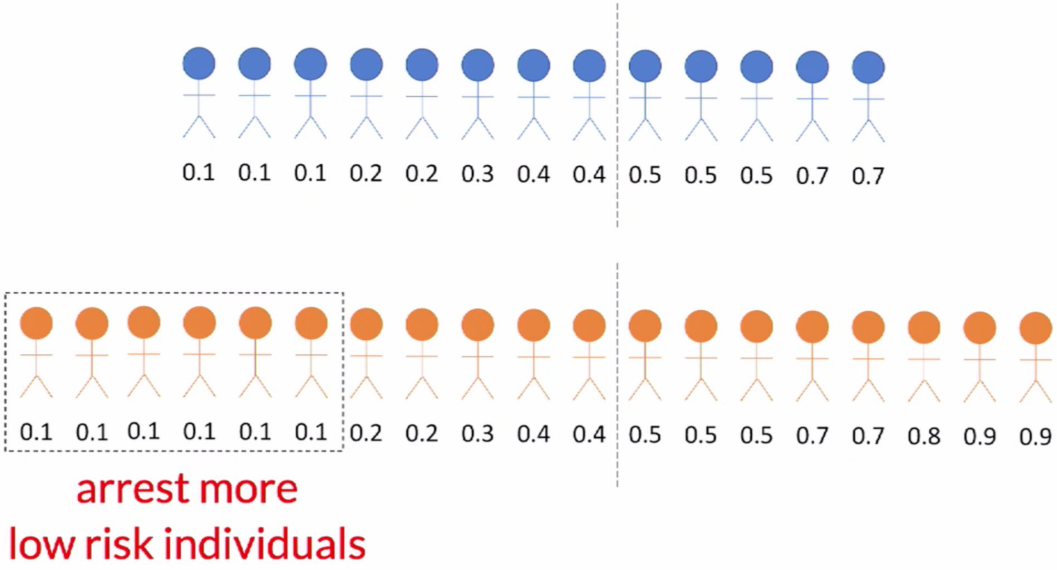
Source: Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks.



Source: A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

Tradeoffs in Fairness

Achieving balance between positive (or negative) class results in a loss in calibration



Source of Image: NIPS’17 Tutorial on Fairness in Machine Learning.

When Can We Achieve Fairness

Two special cases:

- **Perfect Prediction:** can we ever learn a perfect classifier?

- **Non-informative Prediction:** same prediction across the board; tells me nothing!

Inherent trade-offs in the fair determination of risk scores. Kleinberg et al.

Proposed Metrics for Fairness

Notions of fairness and unfairness talked about earlier:

- Calibration within groups
- Balance for the positive class
- Balance for the negative class

- Lack of disparate treatment
- Lack of disparate impact
- Lack of disparate mistreatment

Inherent trade-offs in the fair determination of risk scores.
Kleinberg et al.

Classification without Disparate Mistreatment.
Zafar et al.

Discussion #3

Designing a fair classification algorithm

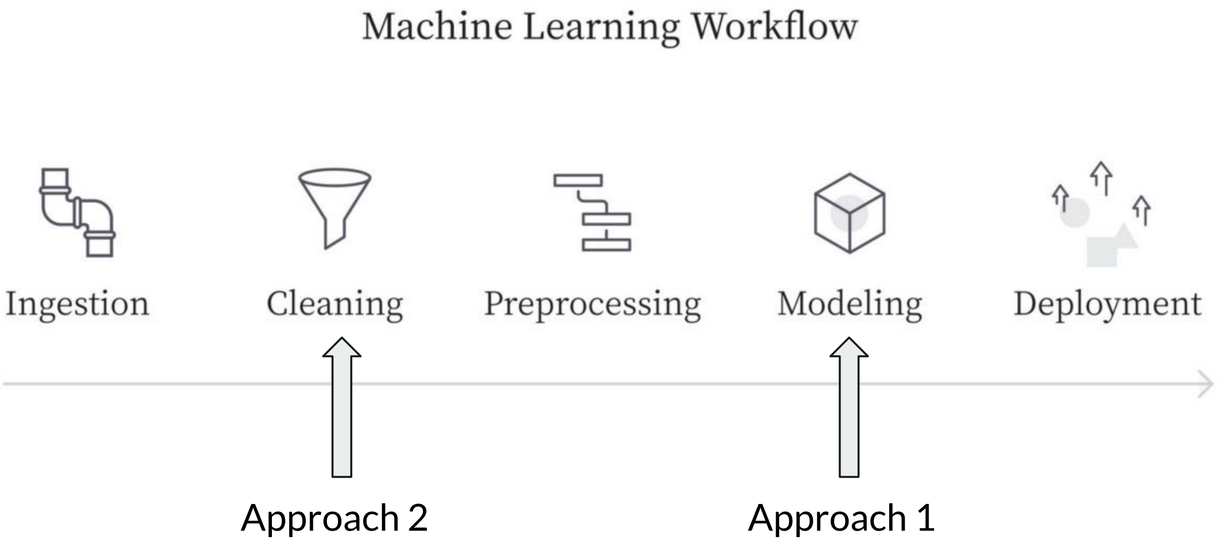
Q1: Given the different notions of fairness; how would you design a fair classifier in risk assessment such as COMPAS? What design decisions would you make or emphasize?

Q2: How would you balance the aforementioned metrics in your model?

Designing A *Fair* & Intelligent System

How to design a "fair" classifier that avoids disparate mistreatment, disparate treatment and balances the misclassification rates across the positive and negative classes

Approaches To Fairness In Systems For Social Use



Source of Image: https://towardsdatascience.com/pipelines-automated-machine-learning-with-hyperparameter-tuning-part-1-b9c06a99d3c3

Approaches To Fairness In Systems For Social Use

Two approaches to designing fair systems

- Algorithmic approaches **(Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment)**
- Data cleaning & preprocessing **(Unequal Representation and Gender Stereotypes in Image Search Results for Occupations)**

Approach 1: Regularization

Empirical Risk Minimization

$$\text{minimize} \quad L(\boldsymbol{\theta})$$

Empirical Risk Minimization Without Disparate Mistreatment

$$\begin{array}{ll} \text{minimize} & L(\boldsymbol{\theta}) \\ \text{subject to} & P(\hat{y} \neq y|z=0) - P(\hat{y} \neq y|z=1) \leq \epsilon, \\ & P(\hat{y} \neq y|z=0) - P(\hat{y} \neq y|z=1) \geq -\epsilon, \end{array} \qquad (8)$$

Bounded difference in the **overall misclassification rate (OMR)** across sensitive groups z

Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. Zafar et al.

Approach 1: Regularization

Approximate this:

$$P(\hat{y} \neq y | z = 0) - P(\hat{y} \neq y | z = 1) \leq \epsilon, \qquad (8)$$
$$P(\hat{y} \neq y | z = 0) - P(\hat{y} \neq y | z = 1) \geq -\epsilon,$$

Potentially non-convex

By:

$$\begin{aligned} \text{Cov}(z, g_{\boldsymbol{\theta}}(y, \mathbf{x})) &= \mathbb{E}[(z - \bar{z})(g_{\boldsymbol{\theta}}(y, \mathbf{x}) - \bar{g}_{\boldsymbol{\theta}}(y, \mathbf{x}))] \\ &\approx \frac{1}{N} \sum_{(\mathbf{x}, y, z) \in \mathcal{D}} (z - \bar{z})\, g_{\boldsymbol{\theta}}(y, \mathbf{x}), \qquad (9) \end{aligned}$$

Is convex

Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. Zafar et al.

Approach 1: Regularization

Original Empirical Risk Minimization Without Disparate Mistreatment

$$\begin{array}{ll} \text{minimize} & L(\boldsymbol{\theta}) \\ \text{subject to} & P(\hat{y} \neq y | z = 0) - P(\hat{y} \neq y | z = 1) \leq \epsilon, \\ & P(\hat{y} \neq y | z = 0) - P(\hat{y} \neq y | z = 1) \geq -\epsilon, \end{array}$$

Proxy Empirical Risk Minimization Without Disparate Mistreatment

$$\begin{array}{ll} \text{minimize} & L(\boldsymbol{\theta}) \\ \text{subject to} & \frac{1}{N} \sum_{(\mathbf{x}, y, z) \in \mathcal{D}} (z - \bar{z})\, g_{\boldsymbol{\theta}}(y, \mathbf{x}) \leq c, \\ & \frac{1}{N} \sum_{(\mathbf{x}, y, z) \in \mathcal{D}} (z - \bar{z})\, g_{\boldsymbol{\theta}}(y, \mathbf{x}) \geq -c, \end{array}$$

Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. Zafar et al.

Approach 1: Regularization

Case study with Logistic Regression:

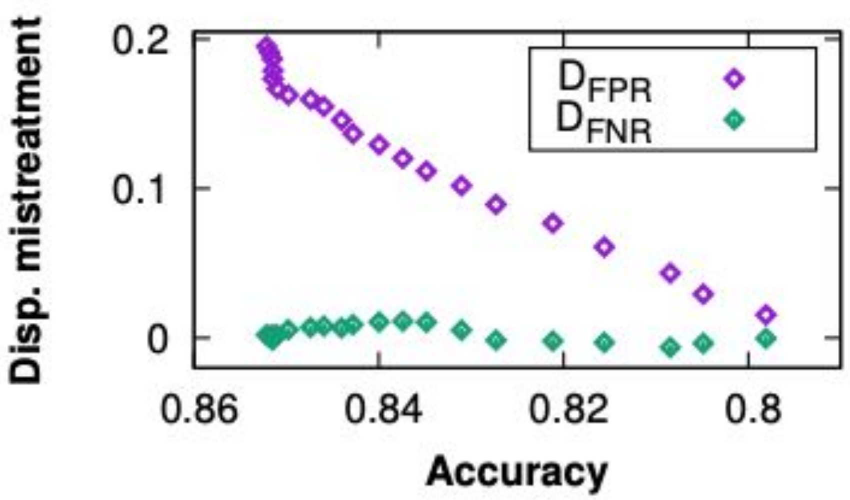
- Disparate mistreatment:
    - OMR vs FPR & FNR

		Predicted Label		
		$\hat{y} = 1$	$\hat{y} = -1$	
True Label	$y = 1$	True positive	False negative	$P(\hat{y} \neq y \mid y = 1)$ False Negative Rate
	$y = -1$	False positive	True negative	$P(\hat{y} \neq y \mid y = -1)$ False Positive Rate
		$P(\hat{y} \neq y \mid \hat{y} = 1)$ False Discovery Rate	$P(\hat{y} \neq y \mid \hat{y} = -1)$ False Omission Rate	$P(\hat{y} \neq y)$ Overall Misclass. Rate

Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. Zafar et al.

Approach 1: Regularization

How does fairness affect accuracy and generalization:



FPR = False Positive Rate

FNR = False Negative Rate

Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. Zafar et al.

Approach 2: Dataset Preprocessing

Data Preprocessing Questions:

- How does prevalence of sensitive features in the dataset correspond to their prevalence in the actual distribution? Are some sensitive features systematically over- or under-represented across domains, and is there stereotype exaggeration in proportions?

- Are there qualitative differences in how the different groups possessing the sensitive features are portrayed in the data generating distribution?

- Do models trained on biased data perpetuate further biases? Are there systemic over- or under-representations of different sensitive groups in the data?

Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. Kay et. al.

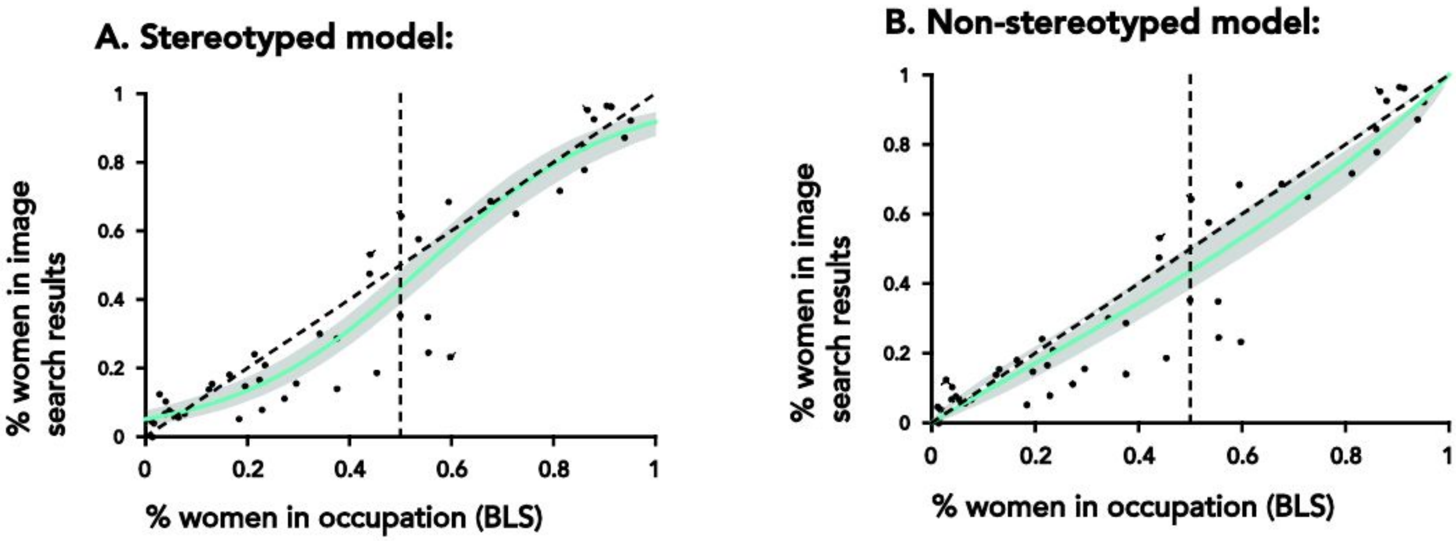
Approach 2: Image Search Dataset Preprocessing

- Sensitive attribute: Gender and their representation in search results

- Filtered image search dataset (with Amazon Turkers) to match "true" population distribution

Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. Kay et. al.

Approach 2: Image Search Dataset Preprocessing



Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. Kay et. al.

Discussion #4

AI and Society

Q1: To what extent should AI and Society interact in sensitive disciplines such as resource allocation, criminal justice, etc…

Conclusion

Thank you!