A Unified Approach to Quantifying Algorithmic Unfairness:

Measuring Individual & Group Unfairness via Inequality Indices

Till Speicher

joint work with

Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi,
Adish Singla, Adrian Weller, Muhammad Bilal Zafar

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

ETH zürich

UNIVERSITY OF CAMBRIDGE

The Alan Turing Institute

Algorithmic Decision Making

Algorithms **assist** and **automate** human decision making

Decisions have **social implications**

Potential for Unfairness

COMPAS: Recidivism risk prediction tool

Machine Bias

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

Risk of white defendants underestimated and risk of black defendants overestimated by algorithm

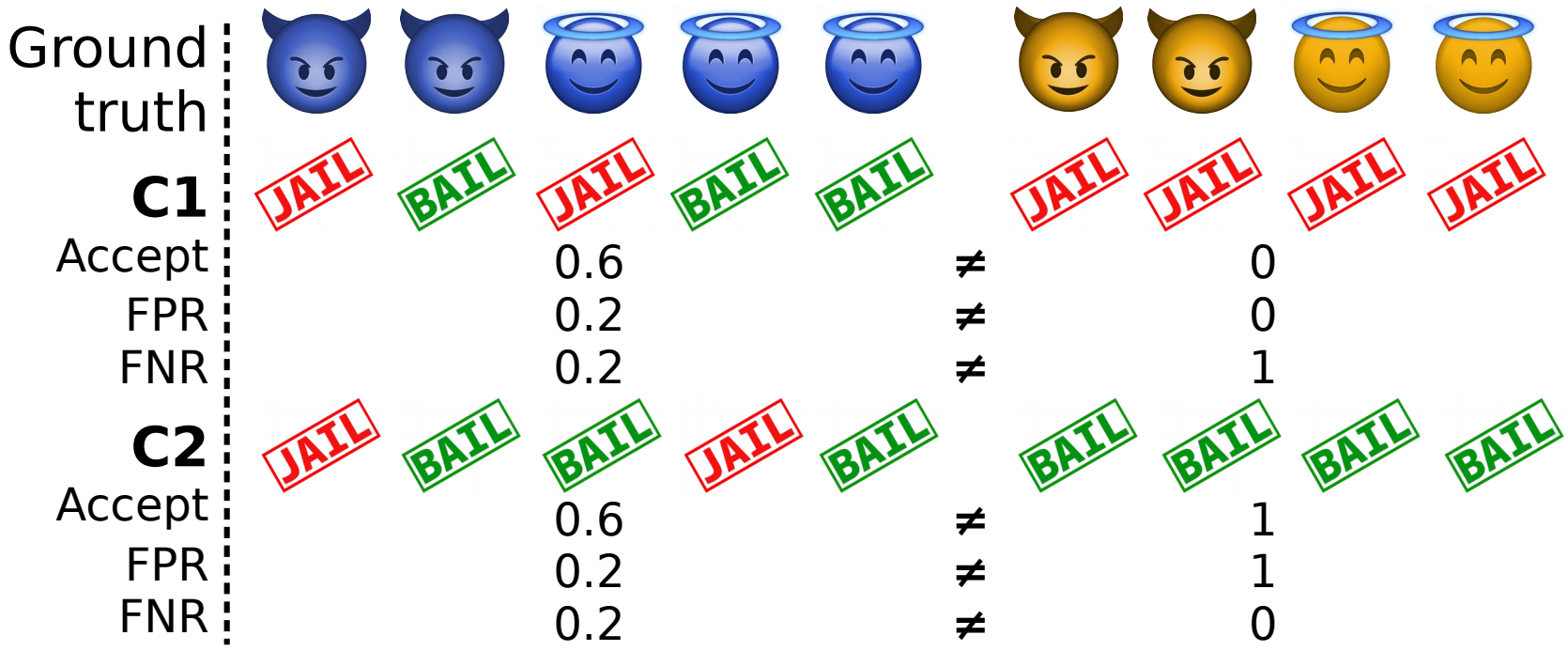
Unfairness in Recidivism Risk Prediction

Are the classifiers fair?

- C1 **biased against** group 2
- C2 **favors** group 2

Which one to choose?

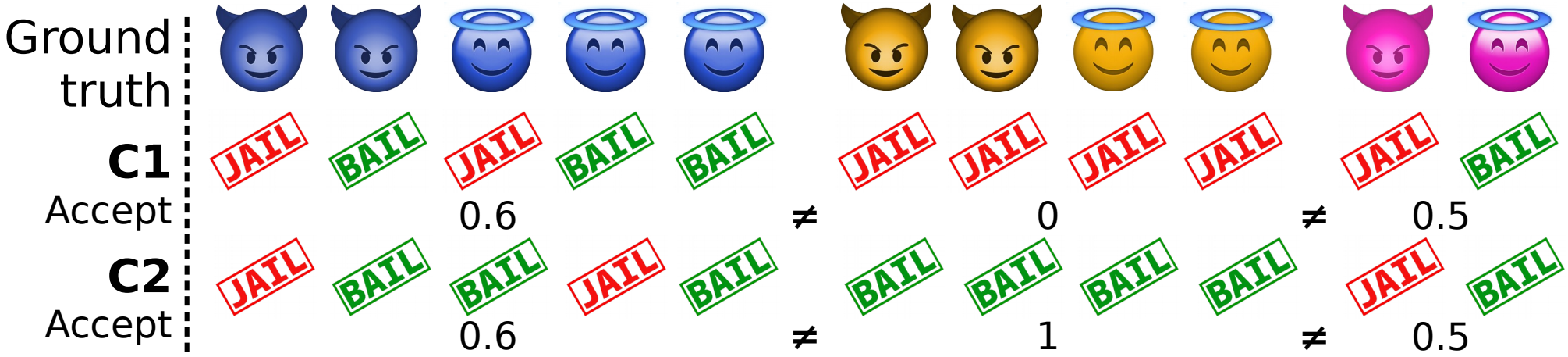
Applying Current Fairness Notions



- ~~Disparate impact/statistical parity:~~
  Equal acceptance rates for each group
- ~~Disparate Mistreatment/equal opportunity:~~
  Equal error rates for each group

Conditions

Current Ways to Measure Unfairness



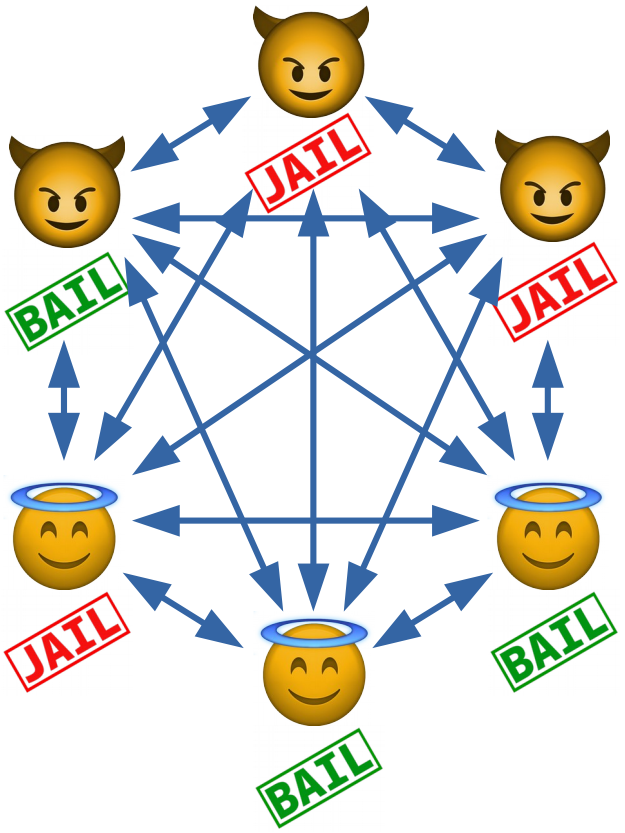
- Most popular measure: **Difference between two group statistics**
- E.g. |Acceptance rate 1
  - Acceptance rate 2|

- Is this a good unfairness measure? What about …
  - … different group sizes?
  - … more than two groups?
  - … non-binary labels?

Individual Fairness

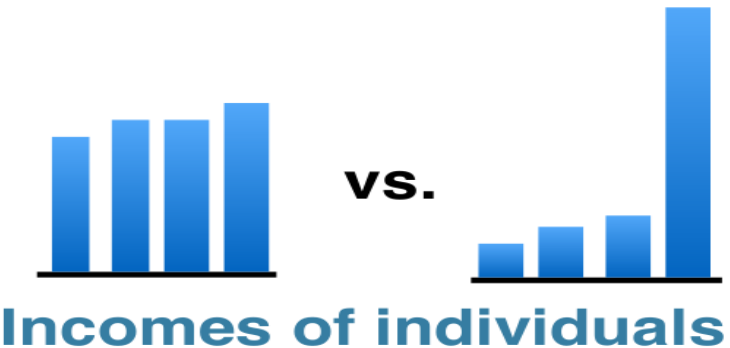
- So far we looked at **group fairness**
- There is also **individual fairness**
- How to measure it?

Need a **principled unfairness measure**



Inspiration: Inequality Indices

- Inequality indices studied in economics
- Measures of **inequality in income distributions** earned by a population
- Principled design



Contributions and Outline

- Define a **principled measure of unfairness** by adapting **inequality indices** to algorithmic decision making
  - Satisfies fairness axioms
  - Adaptable to different types of unfairness
- Reveal relationship between **individual and group fairness**

Inequality Indices

- Many different inequality indices:
  - Gini Index

$$Gini(x_1,\ldots,x_N) = \frac{1}{2N^2\bar{x}}\sum_{i=1}^{N}\sum_{j=1}^{N}|x_i - x_j|$$

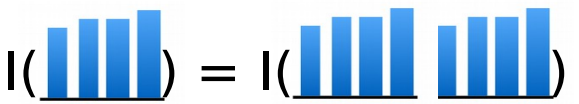
  - Generalized Entropy Indices

$$GE_\alpha(x_1,\ldots,x_N) = \frac{1}{N\alpha(\alpha-1)}\sum_{i=1}^{N}\left[\left(\frac{x_i}{\bar{x}}\right)^\alpha - 1\right] \qquad \alpha \neq 0,1$$

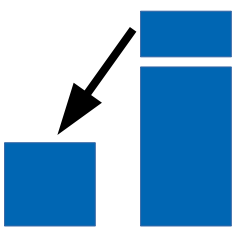
- Designed to satisfy **fairness axioms**

Fairness Axioms

- **Zero-normalization**:
  - Zero inequality if everyone earns the same income
- **Anonymity**:
  - Inequality independent of identity of earners
- **Population invariance**:
  - Metric does not depend on size of population
- **Transfer principle**:
  - Income transfer from high- to low-earning individuals decreases inequality



Inequality decrease



Converting Algorithmic Decisions to Benefits

- Inequality indices designed to measure inequality in incomes
- For application in algorithmic decision making:
  Need to map deserved and predicted outcomes to **benefits**
- Example:

Benefit		Deserved	
		BAIL	JAIL
Predicted	BAIL	**1**	**2**
	JAIL	**0**	**1**

- We show: Suitable benefit functions **capture fairness notions** based on
  - Acceptance rate
  - FPR
  - FNR
  - …

Applying Inequality Indices

Ground truth	😈	😈	😇	😇	😇	😈	😈	😇	😇	😈	😇
C1	**1**	**2**	**0**	**1**	**1**	**1**	**1**	**0**	**0**	**1**	**1**
C2	**1**	**2**	**1**	**0**	**1**	**2**	**2**	**1**	**1**	**1**	**1**

- Generalized Entropy Index (α = 2):

$$GE_2(b_1,\ldots,b_N) = \frac{1}{2N}\sum_{i=1}^{N}\left[\left(\frac{b_i}{\bar{b}}\right)^2 - 1\right]$$

- Inequalities:
  - C1: 0.25
  - C2: 0.12 → less unfair

Individual unfairness

Applying Inequality Indices: Group Fairness

Ground truth	😈	😈	😇	😇	😇	😈	😈	😇	😇	😈	😇
C1	**1**	**1**	**1**	**1**	**1**	**0.5**	**0.5**	**0.5**	**0.5**	**1**	**1**
C2	**1**	**1**	**1**	**1**	**1**	**1.5**	**1.5**	**1.5**	**1.5**	**1**	**1**

- Replacing individual benefits with groups' mean benefits (b')
- Generalized Entropy, **between-group** component:

$$GE_{between}(b'_1,\ldots,b'_N) = \frac{1}{2N}\sum_{i=1}^{N}\left[\left(\frac{b'_i}{\bar{b}}\right)^2 - 1\right]$$

- Between-group inequalities:
  - C1: 0.04
  - C2: 0.02 → less group-unfair

Contributions and Outline

- Define a **principled measure of unfairness** using **inequality indices**
  - Satisfies fairness axioms
  - Adaptable to different types of fairness
- Reveal relationship between **individual and group fairness**

Connecting Individual and Group Fairness

- The first solution was measuring unfairness **between individuals** instead of **groups**
- Some inequality indices are **subgroup decomposable**:

  Overall (individual) inequality is the sum of inequality $I(b) =$
  - **Between** (means of) subgroups $I_\beta(b) +$
  - **Within** each subgroup $I_\omega(b)$
- Not satisfied by all indices, **Generalized Entropy** family does

Decomposing Unfairness

- SD allows **decomposition** of

Overall unfairness into Between-group unfairness + Within-group unfairness

$$\mathcal{E}^{\alpha}(b_1, b_2, \cdots, b_n) = \sum_{g=1}^{|G|} \frac{n_g}{n\alpha(\alpha-1)} \left[\left(\frac{\mu_g}{\mu} \right)^{\alpha} - 1 \right] + \sum_{g=1}^{|G|} \frac{n_g}{n} \left(\frac{\mu_g}{\mu} \right)^{\alpha} \mathcal{E}^{\alpha}(\mathbf{b}^g)$$

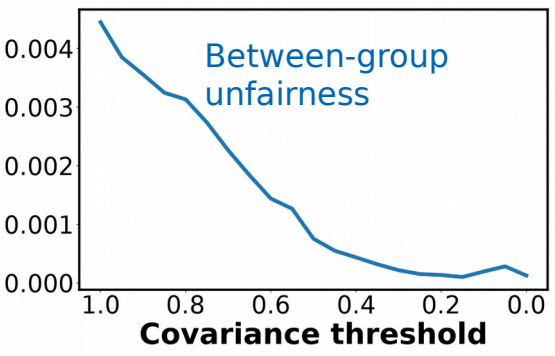
I (😈😈😇😇😇 😈😈😇😇 😈😇)

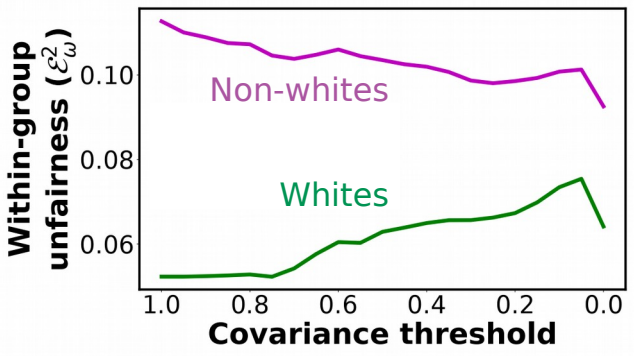
$= I_{\beta}$(μ_1 μ_1 μ_1 μ_1 μ_1 μ_2 μ_2 μ_2 μ_2 μ_3 μ_3)

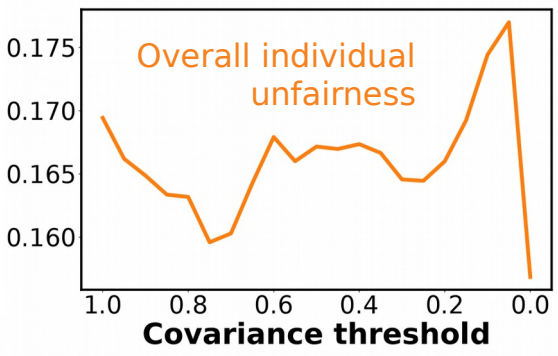
$+ I_{\omega}$(😈😈😇😇😇) $+ I_{\omega}$(😈😈😇😇) $+ I_{\omega}$(😈😇)

Fairness Tradeoffs via Decomposition

- Prior work on unfairness in machine learning:
  Focussed on detecting and eliminating **discrimination**
- Ignores **fairness tradeoffs**







Eliminating between-group unfairness **can increase** within-group or overall individual **unfairness**

Summary

- Introduce Inequality indices as a **principled** measure of algorithmic unfairness
- Take a **unified approach** to measuring unfairness where **overall individual** unfairness is decomposed into **between**- and **within-group** unfairness
- Future work: Training models to **eliminate** overall individual and within-group **unfairness**