Understanding and Accounting for Human Perceptions of Fairness in Algorithmic Decision Making

Nina Grgić-Hlača, Elissa M. Redmiles, Muhammad Bilal Zafar, Krishna P. Gummadi and Adrian Weller

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

UNIVERSITY OF MARYLAND

UNIVERSITY OF CAMBRIDGE

Machine-assisted Decision Making

Algorithms help people make decisions

Hiring

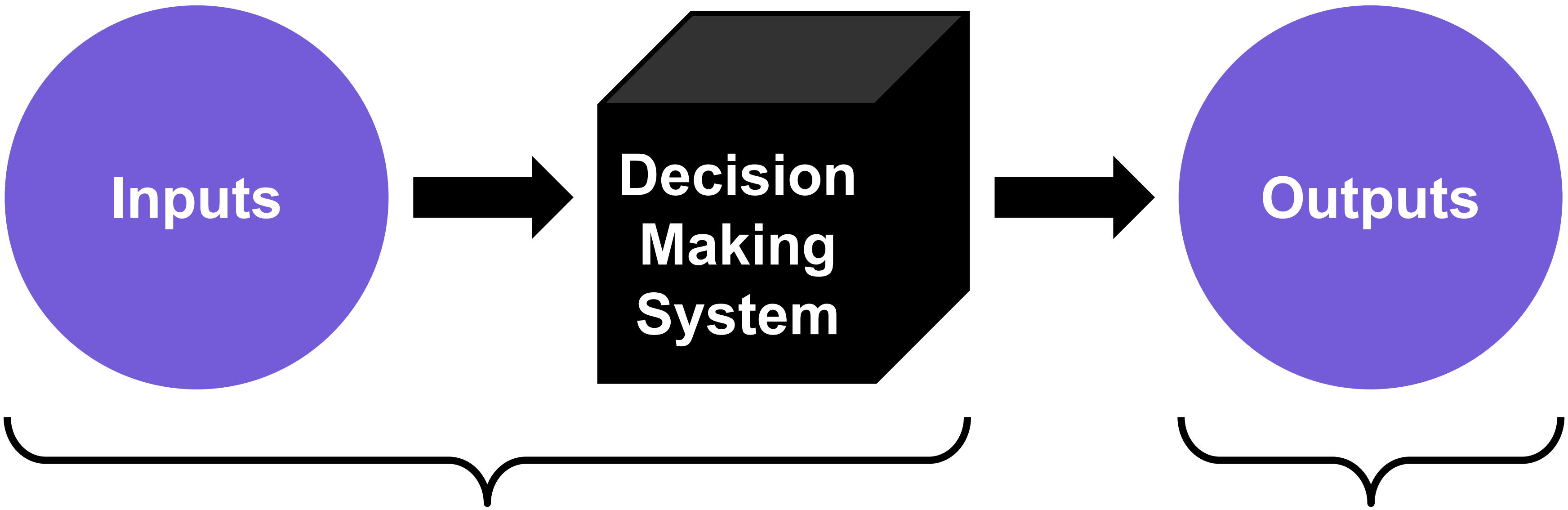
Social benefits

Granting bail

Are these algorithms **fair**?

Decision Making Pipeline

Example: **Granting bail**



Is it fair to use a feature?

Equal **error rates**?

Is it Fair to Use a Feature?

Normative approach

Prescribe how fair decisions ought to be made

Anti-discrimination laws

- **Sensitive** (**race**, **gender**) vs **non-sensitive** features

Descriptive approach

Describe human perceptions of fairness

Beyond discrimination?

- Father's criminal history
- Education
- Ice-cream preference

This Talk

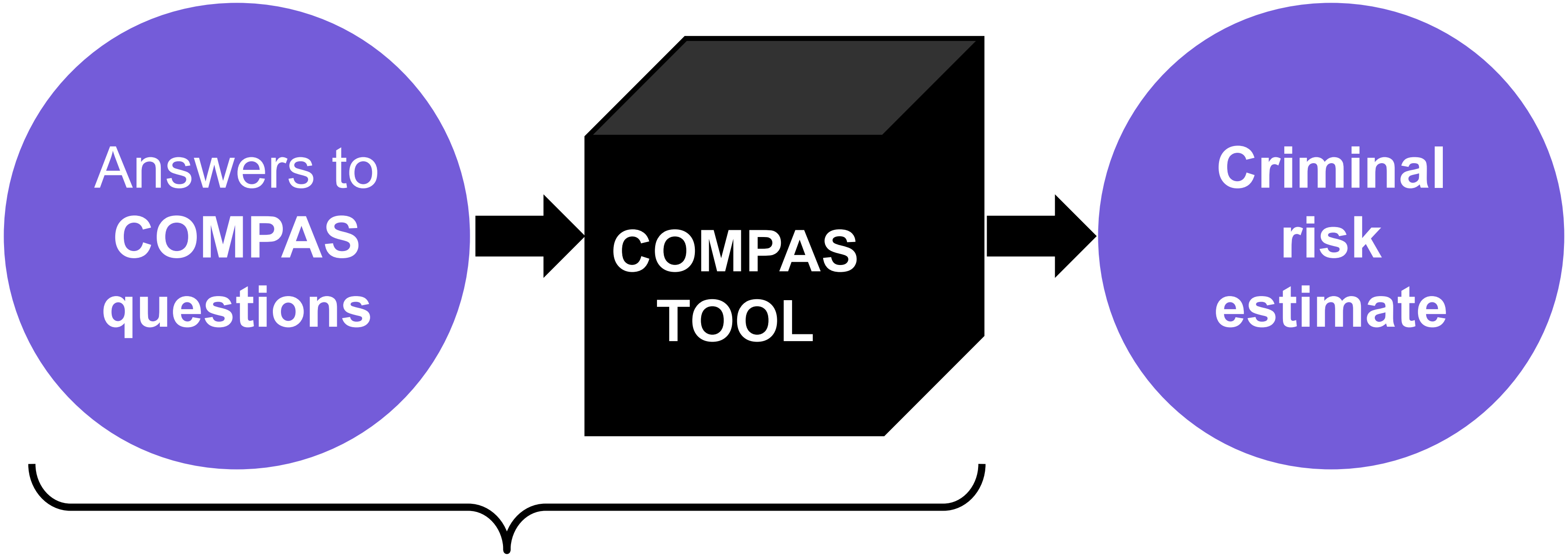
- **Which** **features people perceive as** **fair** **to use?**
- **Why** **do people** **perceive** **some features as unfair?**
- **How** **to** **account for** **people's fairness perceptions?**

This Talk

- **Which features people perceive as fair to use?**
- **Why do people perceive some features as unfair?**
- **How to account for people's fairness perceptions?**

Assisting Bail Decisions

Case Study: **COMPAS Tool**



Which **features**?

COMPAS Questionnaire

137 questions, 10 topics

Current criminal charges	Criminal attitudes
Criminal history	Neighborhood safety
Substance abuse	Criminal history of friends & family
Stability of employment	Quality of social life
Personality	Education & behavior in school

No questions about **sensitive features!**

Is it **fair** to use these features to **make bail decisions**?

Gathering Human Moral Judgments

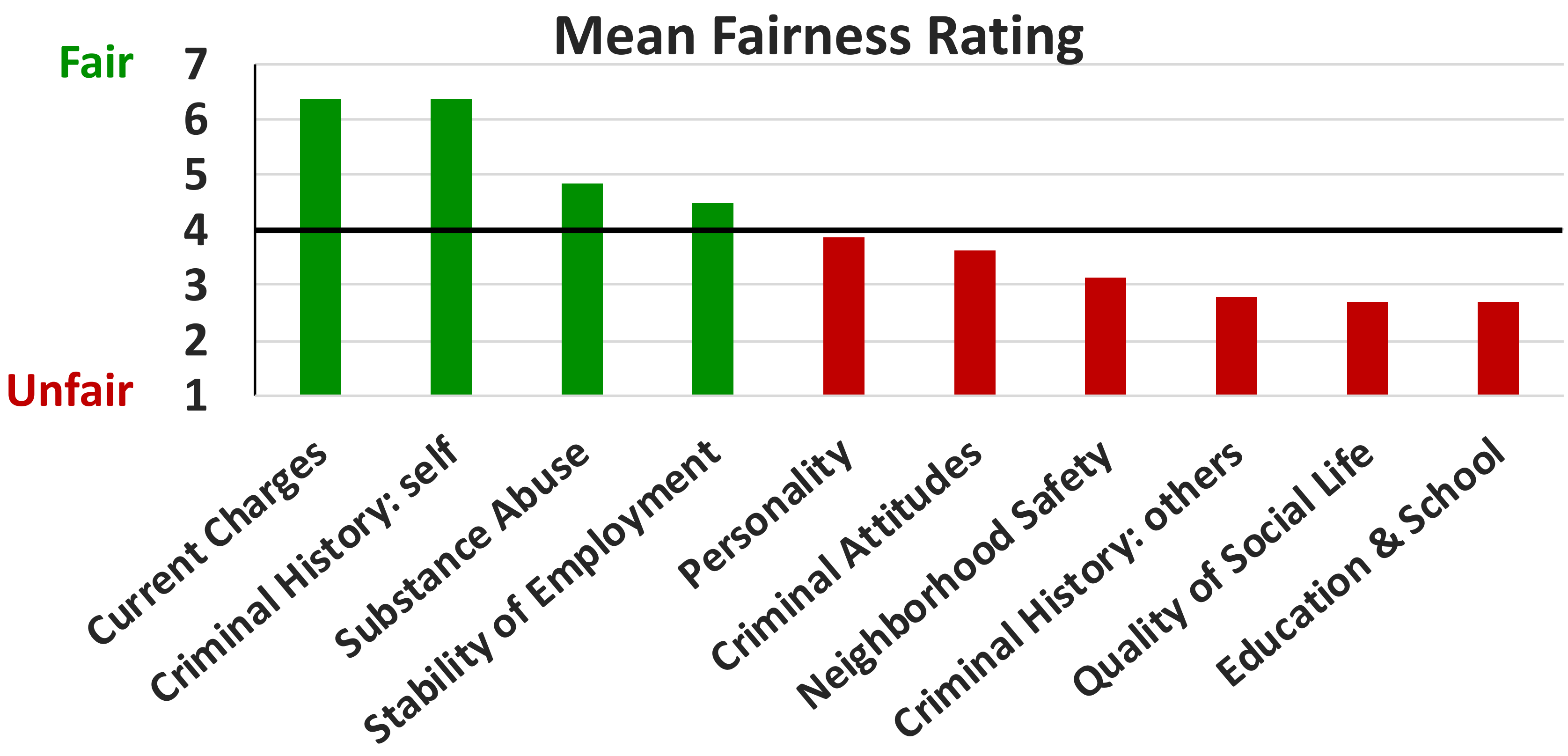
- **Fairness of using features** for making bail decisions

- US criminal justice system – **US respondents**
  - 196 **Amazon Mechanical Turk master** workers
  - 380 **SSI** survey panel respondents, census representative

Findings **consistent** across both samples

Is it Fair to Use these Features?

People consider **most** of the features **unfair**!



This Talk

- **Which features people perceive as fair to use?**
- **Why do people perceive some features as unfair?**
- **How to account for people's fairness perceptions?**

This Talk

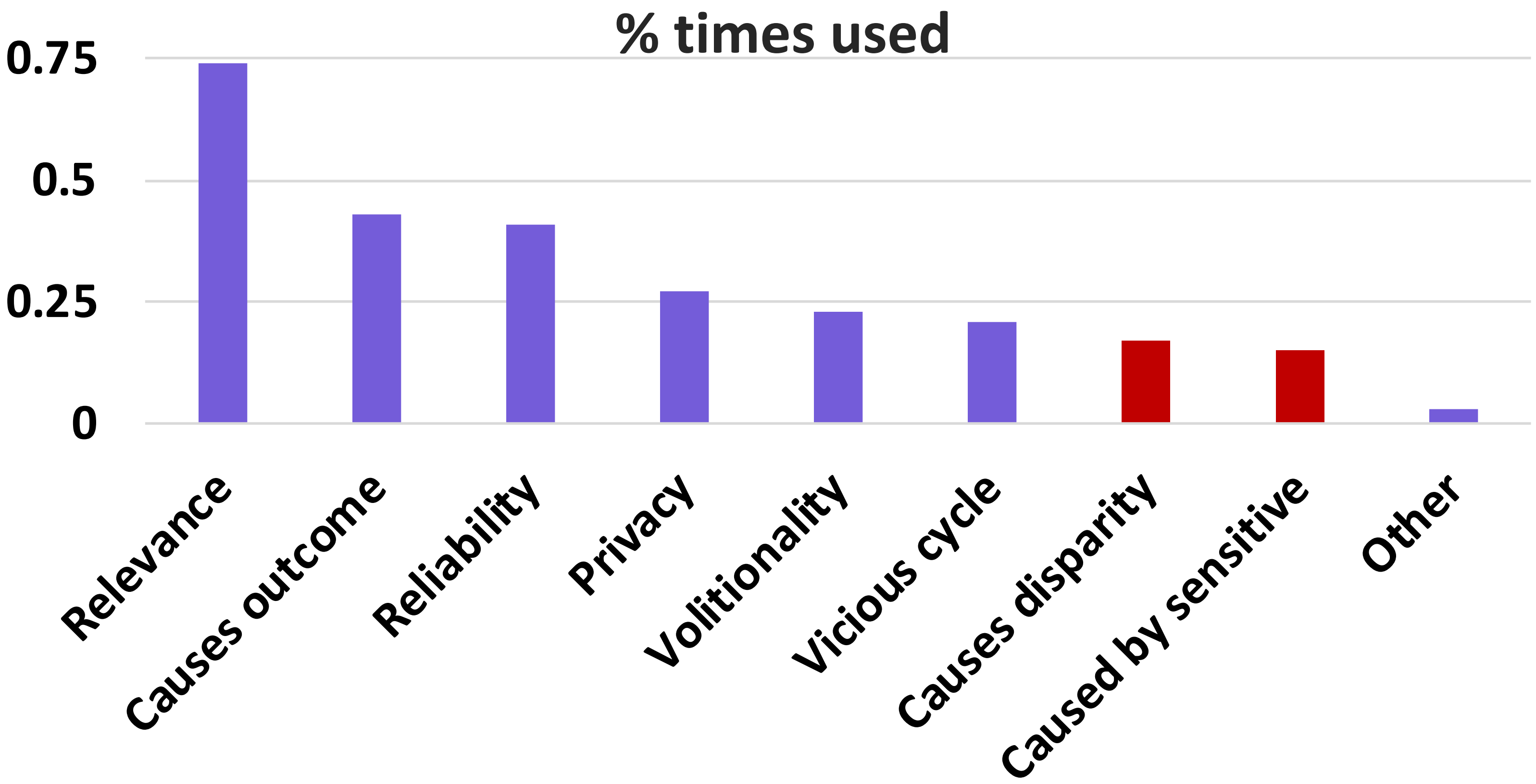
- Which features people perceive as fair to use?
- **Why do people perceive some features as unfair?**
- How to account for people's fairness perceptions?

Hypothesis I:
Latent Properties of Features

Relevant?
Reliable?
Volitional?
Private?
Causes Outcome?
Causes Vicious Cycle?
Causes Disparity in Outcomes?
Caused by Sensitive Group Membership?

What Makes a Feature (un)Fair to Use?

There is **more to fairness than discrimination!**

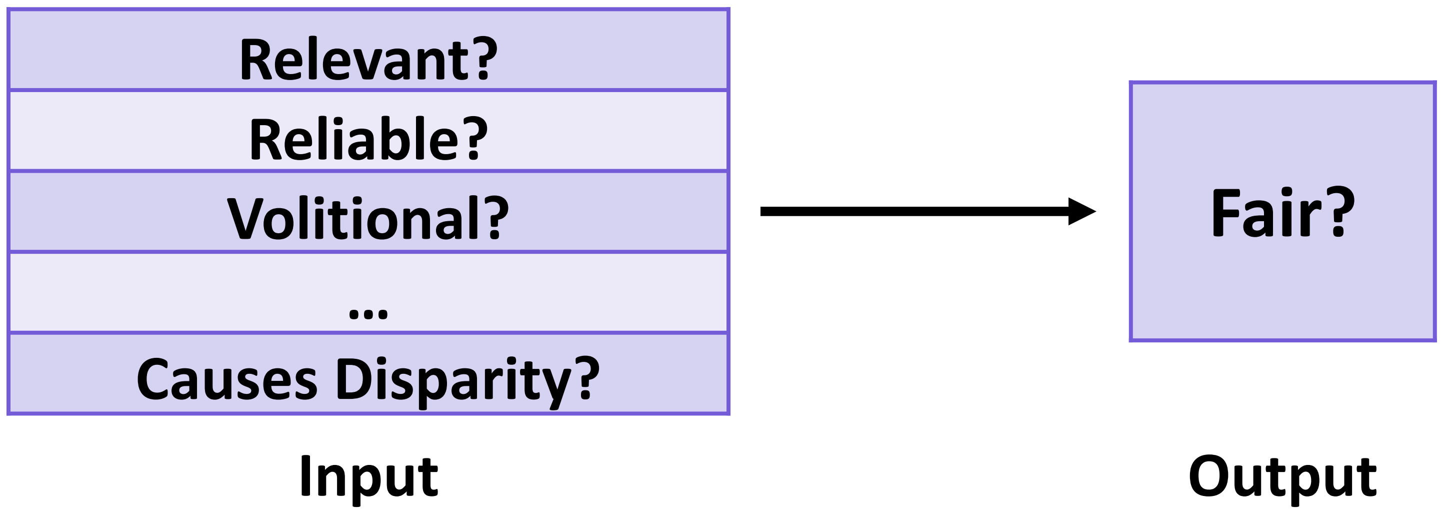


Hypothesis II: From Latent Properties to Fairness

Relevant?
Reliable?
Volitional?
Private?
Causes Outcome?
Causes Vicious Cycle?
Causes Disparity in Outcomes?
Caused by Sensitive Group Membership?

→ **Fair?**

Modeling Fairness Judgments



We can predict fairness judgments with **88% accuracy**

We model a **common fairness judgment heuristic**

- May be culturally dependent: interesting future work

This Talk

- Which features people perceive as fair to use?
- **Why do people perceive some features as unfair?**
- How to account for people's fairness perceptions?

Take-aways

Q: Is it **fair** to use a feature?

A: Depends on the feature's **latent properties**!

- **Relevance**
- **Reliability**
- **Volitionality**
- **Privacy**
- **Causal relationships**

Fairness beyond **discrimination**

This Talk

- Which features people perceive as fair to use?
- Why do people perceive some features as unfair?
- How to account for people's fairness perceptions?

Accounting for Fairness Judgments

Goal: Train machine learning algorithms that

- Achieve high **accuracy**
- People **perceive as fair**

Prerequisite: **measure** these quantities

- We know how to measure accuracy
- How do we **measure perceived fairness**?

Quantifying Perceived Fairness

Fairness of using a feature

- **Fraction** of people that consider using the feature fair

Fairness of using classifier

- Fraction of people that consider **all** of its **features fair**

Accounting for Fairness Judgments

Goal: Train machine learning algorithms that

- Achieve high **accuracy**
- People **perceive as fair**

Implement: Select subset of **features** that

$$\begin{aligned} \underset{S \subseteq \mathcal{F}}{\text{maximize}} \quad & accuracy(\mathcal{S}) \\ \text{subject to} \quad & unfairness(\mathcal{S}) \leq t \end{aligned}$$

Perceived Fairness vs Accuracy

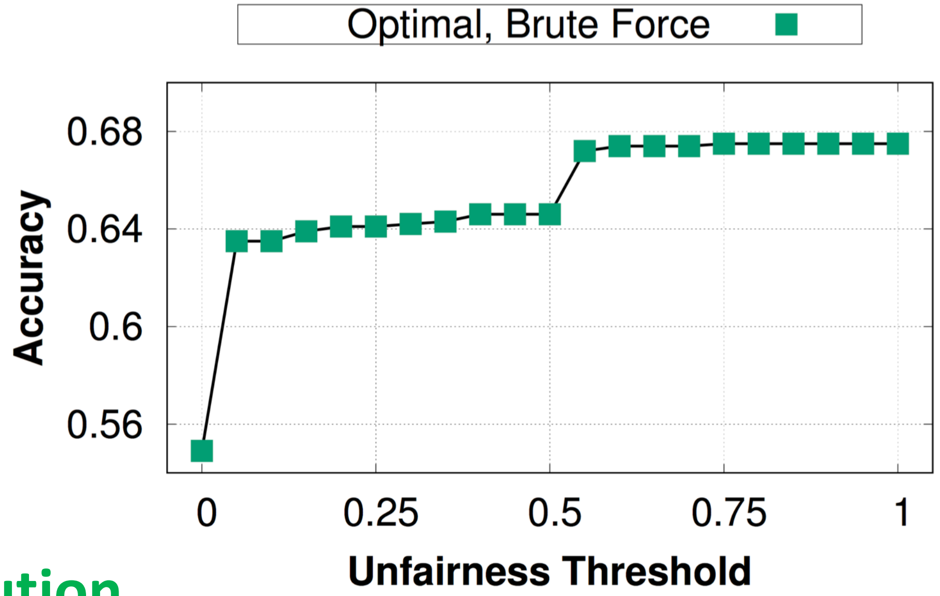
Intuition

- Adding features: higher accuracy, lower fairness
- Removing features: lower accuracy, higher fairness

There is a **tradeoff** between **perceived fairness of features** & **accuracy**

Naïve Approach

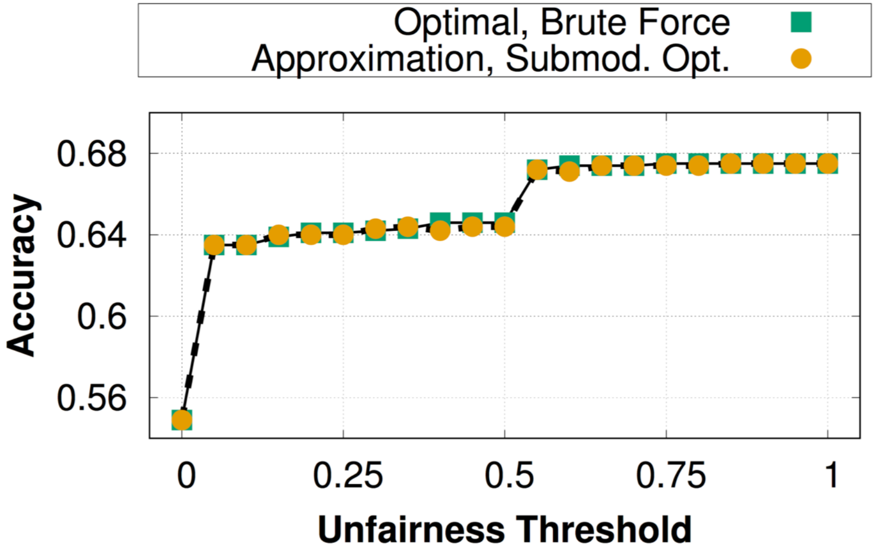
- **Brute force**
  - **Train 2^n classifiers**, n = number of features



- **Optimal Solution**
- **Not scalable!** 30 features = more than 1 billion classifiers
- **Is there an efficient alternative?**

Submodular Optimization

- **Feature usage unfairness** is **submodular** & **monotone**
- **Submodular cost submodular knapsack problem**
  - Approximate using **ISK** algorithm (Iyer and Bilmes, NIPS 2013)

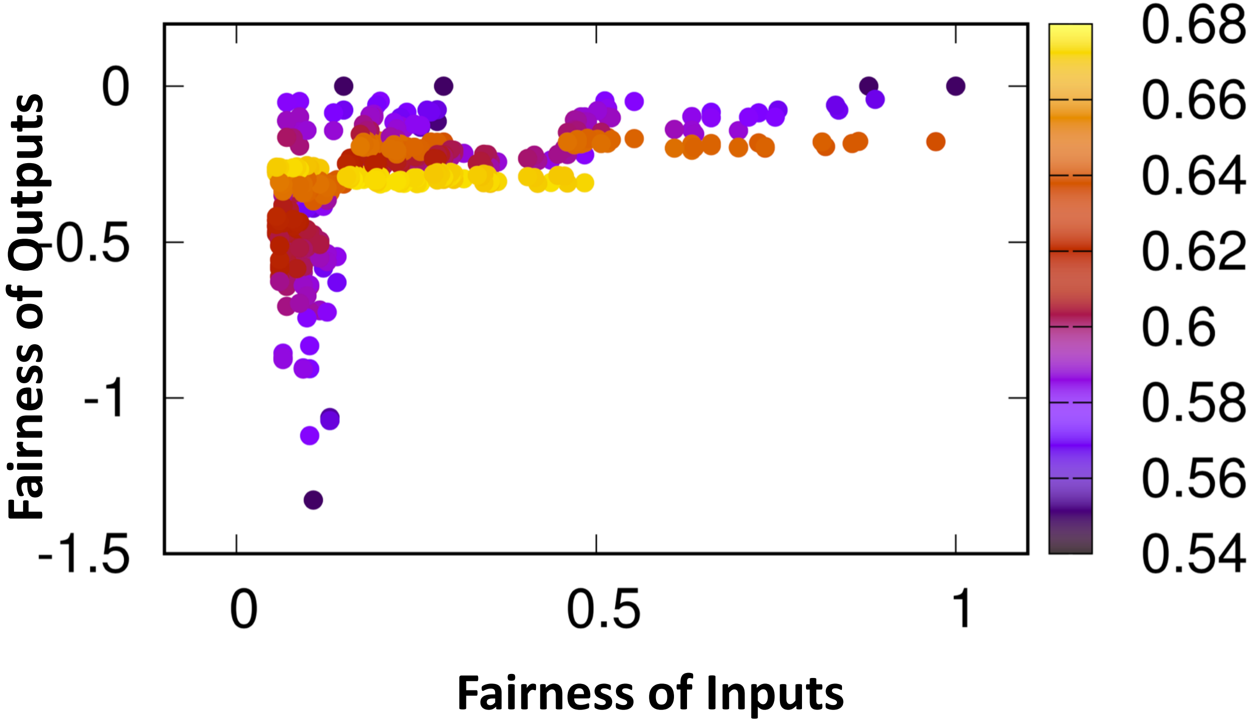


- **Efficient & scalable approximation**
- **Near optimal results**

Fair Inputs vs Fair Outputs

- Fairness of outputs: equal misclassification rates
- In the **ProPublica COMPAS dataset**:

Fair inputs → fair outputs



Take-aways

Understanding Human Perceptions of Fairness

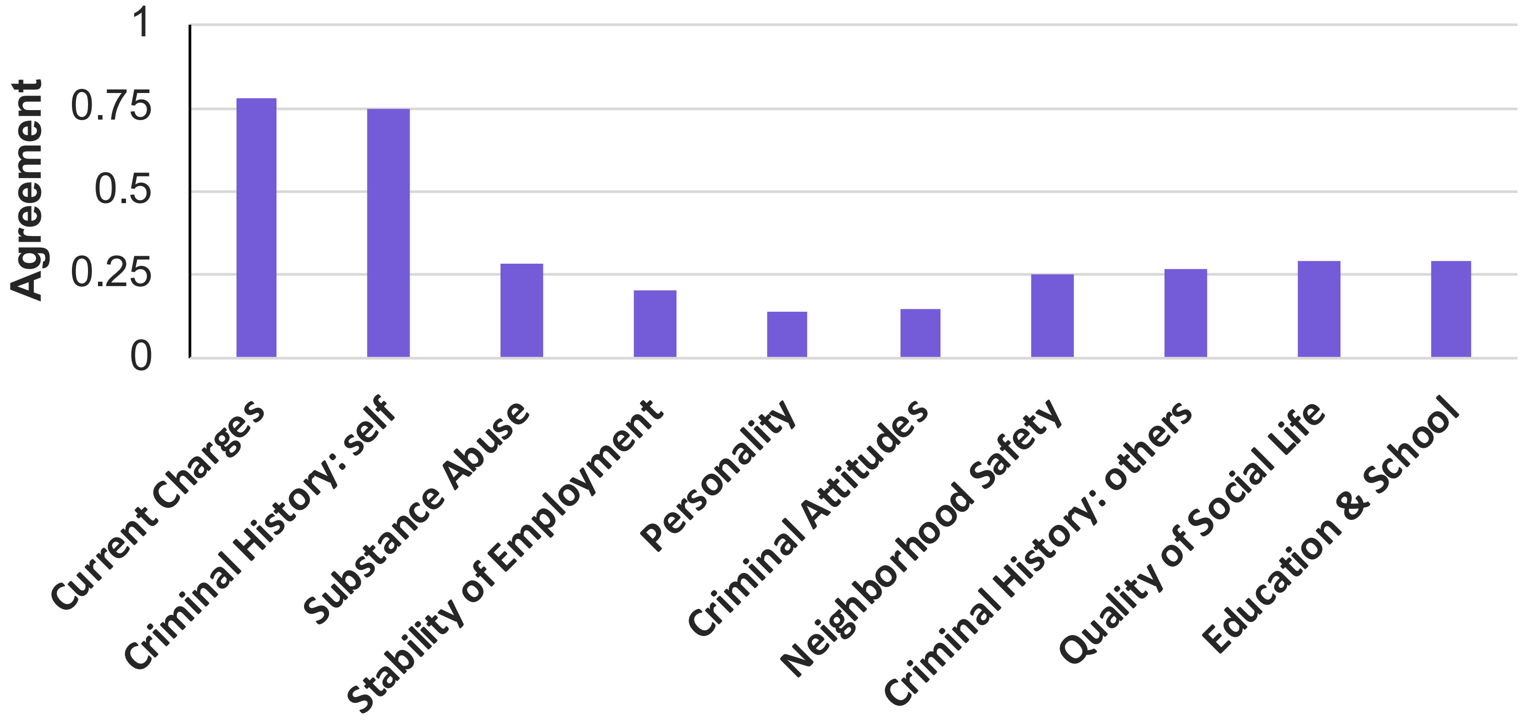
- From **latent properties** to **fairness judgments**
- Fairness considerations go **beyond discrimination**

Accounting for Human Perceptions of Fairness

- **Measure** that captures perceptions of feature usage fairness
- **Mechanism** for selecting features perceived as fair

Bonus Slides - Understanding

Do people agree in their fairness judgments?



People often **disagree** in their fairness judgments

Causes of Disagreements in Fairness Judgments

How can we explain **disagreements** in fairness judgments?

Relevant?
Reliable?
Volitional?
…
Causes Disparity?

Properties of a Feature
Disagreement?

→

Mapping
Agreement

Fair?

Fairness of Using a Feature
Disagreement

Disagreements in Latent Property Assessments?

Relevant?
Reliable?
Private?
Volitional?
Causes Outcome?
Causes Vicious Cycle?
Causes Disparity in Outcomes?
Caused by Sensitive Group Membership?

Some consensus

Low consensus

Causal properties

Constructing causal graphs?

Causes of Disagreements in Fairness Judgments

How can we explain **disagreements** in fairness judgments?

Relevant?
Reliable?
Volitional?
…
Causes Disparity?

→

Fair?

Properties of a Feature
~~Disagreement?~~
Disagreement!

Mapping
Agreement

Fairness of Using a Feature
Disagreement

Bonus Slides - Accounting

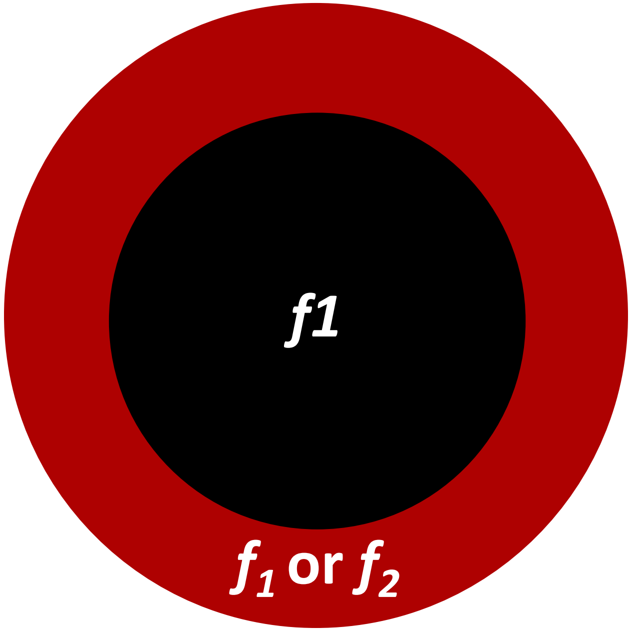
Fairness Properties - Monotonicity

- Feature unfairness is **monotone non-decreasing**

- *Intuition*
  - A set function is monotone non-decreasing if **adding elements to a set cannot decrease its value**

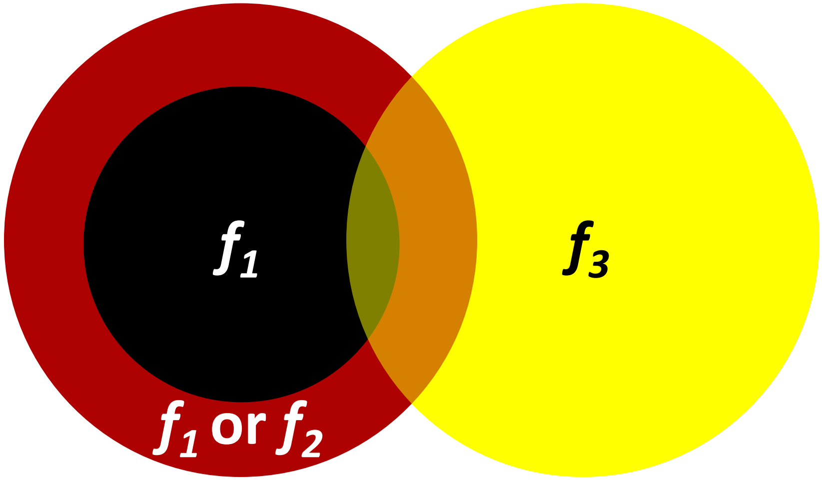
- *Definition*

$$g(\mathcal{F}_i \cup \{f\}) \geq g(\mathcal{F}_i), \quad \forall \mathcal{F}_i \subseteq \mathcal{F}, f \in \mathcal{F} \setminus \mathcal{F}_i$$



Fairness Properties - Submodularity

- Feature unfairness is **submodular**
- *Intuition*
  - A set function is submodular if it exhibits **diminishing marginal returns**



- *Definition*

$$g(\mathcal{F}_A \cup \{f\}) - g(\mathcal{F}_A) \geq g(\mathcal{F}_B \cup \{f\}) - g(\mathcal{F}_B),$$
$$\mathcal{F}_A \subseteq \mathcal{F}_B \subset \mathcal{F}, f \in \mathcal{F} \setminus \mathcal{F}_B$$

ISK algorithm

Problem

$$\begin{aligned} \underset{S \subseteq \mathcal{F}}{\text{maximize}} \quad & accuracy(\mathcal{S}) \\ \text{subject to} \quad & unfairness(\mathcal{S}) \leq t \end{aligned}$$

- Maps to **Submodular Cost Submodular Knapsack** problem

Algorithm – Intuition

- Iteratively **finding modular approximations of submodular functions**
- Solving the resulting knapsack problems