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**

- **Inputs**
- **Decision Making System**
- **Outputs**

- 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

Answers to COMPAS questions

COMPAS TOOL

Criminal risk estimate

Which features?
<table>
<thead>
<tr>
<th>Current criminal charges</th>
<th>Criminal attitudes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criminal history</td>
<td>Neighborhood safety</td>
</tr>
<tr>
<td>Substance abuse</td>
<td>Criminal history of friends &amp; family</td>
</tr>
<tr>
<td>Stability of employment</td>
<td>Quality of social life</td>
</tr>
<tr>
<td>Personality</td>
<td>Education &amp; behavior in school</td>
</tr>
</tbody>
</table>

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**!

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean Fairness Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current Charges</td>
<td>7</td>
</tr>
<tr>
<td>Criminal History: self</td>
<td>6</td>
</tr>
<tr>
<td>Substance Abuse</td>
<td>5</td>
</tr>
<tr>
<td>Stability of Employment</td>
<td>4</td>
</tr>
<tr>
<td>Personality</td>
<td></td>
</tr>
<tr>
<td>Criminal Attitudes</td>
<td>3</td>
</tr>
<tr>
<td>Neighborhood Safety</td>
<td>2</td>
</tr>
<tr>
<td>Criminal History: others</td>
<td>1</td>
</tr>
<tr>
<td>Quality of Social Life</td>
<td>1</td>
</tr>
<tr>
<td>Education &amp; School</td>
<td>1</td>
</tr>
</tbody>
</table>
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?
<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant?</td>
</tr>
<tr>
<td>Reliable?</td>
</tr>
<tr>
<td>Volitional?</td>
</tr>
<tr>
<td>Private?</td>
</tr>
<tr>
<td>Causes Outcome?</td>
</tr>
<tr>
<td>Causes Vicious Cycle?</td>
</tr>
<tr>
<td>Causes Disparity in Outcomes?</td>
</tr>
<tr>
<td>Caused by Sensitive Group Membership?</td>
</tr>
</tbody>
</table>
What Makes a Feature (un)Fair to Use?

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

% times used

- Relevance
- Causes outcome
- Reliability
- Privacy
- Volitionality
- Vicious cycle
- Causes disparity
- Caused by sensitive
- Other
Hypothesis II: From Latent Properties to Fairness

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant?</td>
<td></td>
</tr>
<tr>
<td>Reliable?</td>
<td></td>
</tr>
<tr>
<td>Volitional?</td>
<td></td>
</tr>
<tr>
<td>Private?</td>
<td></td>
</tr>
<tr>
<td>Causes Outcome?</td>
<td></td>
</tr>
<tr>
<td>Causes Vicious Cycle?</td>
<td></td>
</tr>
<tr>
<td>Causes Disparity in Outcomes?</td>
<td></td>
</tr>
<tr>
<td>Caused by Sensitive Group Membership?</td>
<td></td>
</tr>
</tbody>
</table>

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

We model a **common fairness judgment heuristic**

- May be culturally dependent: interesting future work
• Which features people perceive as fair to use?

• Why do people perceive some features as unfair?

• How to account for people’s fairness perceptions?
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

\[
\begin{align*}
\text{maximize} & \quad \text{accuracy}(S) \\
\text{subject to} & \quad \text{unfairness}(S) \leq t
\end{align*}
\]
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, $n = \text{number of features}$

- 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 $\rightarrow$ 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
People often **disagree** in their fairness judgments
How can we explain *disagreements* in fairness judgments?

<table>
<thead>
<tr>
<th>Relevant?</th>
<th>Reliable?</th>
<th>Volitional?</th>
<th>Causes Disparity?</th>
</tr>
</thead>
</table>

- Fair?
- Fairness of Using a Feature Disagreement
- Mapping Agreement
- Properties of a Feature Disagreement?
Disagreements in Latent Property Assessments?

Causal properties

Constructing causal graphs?

Relevant?

Reliable?

Private?

Volitional?

Causes Outcome?

Causes Vicious Cycle?

Causes Disparity in Outcomes?

Caused by Sensitive Group Membership?

Some consensus

Low consensus
How can we explain disagreements in fairness judgments?

<table>
<thead>
<tr>
<th>Relevant?</th>
<th>Fair?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliable?</td>
<td></td>
</tr>
<tr>
<td>Volitional?</td>
<td></td>
</tr>
<tr>
<td>Causes Disparity?</td>
<td>Fairness of Using a Feature Disagreement</td>
</tr>
</tbody>
</table>

Properties of a Feature Disagreement?

Mapping Agreement
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*

\[
g(F_i \cup \{f\}) \geq g(F_i),
\]  
\[
\forall F_i \subseteq \mathcal{F}, f \in \mathcal{F} \setminus F_i
\]
Fairness Properties - Submodularity

• Feature unfairness is submodular

• Intuition
  • A set function is submodular if it exhibits diminishing marginal returns

• Definition

\[
g(F_A \cup \{f\}) - g(F_A) \geq g(F_B \cup \{f\}) - g(F_B),
\]

\[
F_A \subseteq F_B \subset F, f \in F \setminus F_B
\]
Problem

\[
\begin{align*}
\max_{S \subseteq \mathcal{F}} & \quad \text{accuracy}(S) \\
\text{subject to} & \quad \text{unfairness}(S) \leq t
\end{align*}
\]

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

Algorithm – Intuition

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