

Cross-Cultural Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction

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



MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS



UNIVERSITY OF
MARYLAND



UNIVERSITY OF
CAMBRIDGE

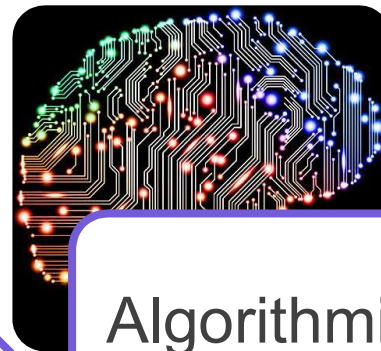
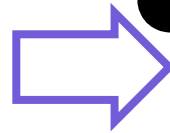


LEVERHULME CENTRE FOR THE
FUTURE OF INTELLIGENCE

Algorithmic Decision Making



Human
decision
making



Algorithmic
decision
making

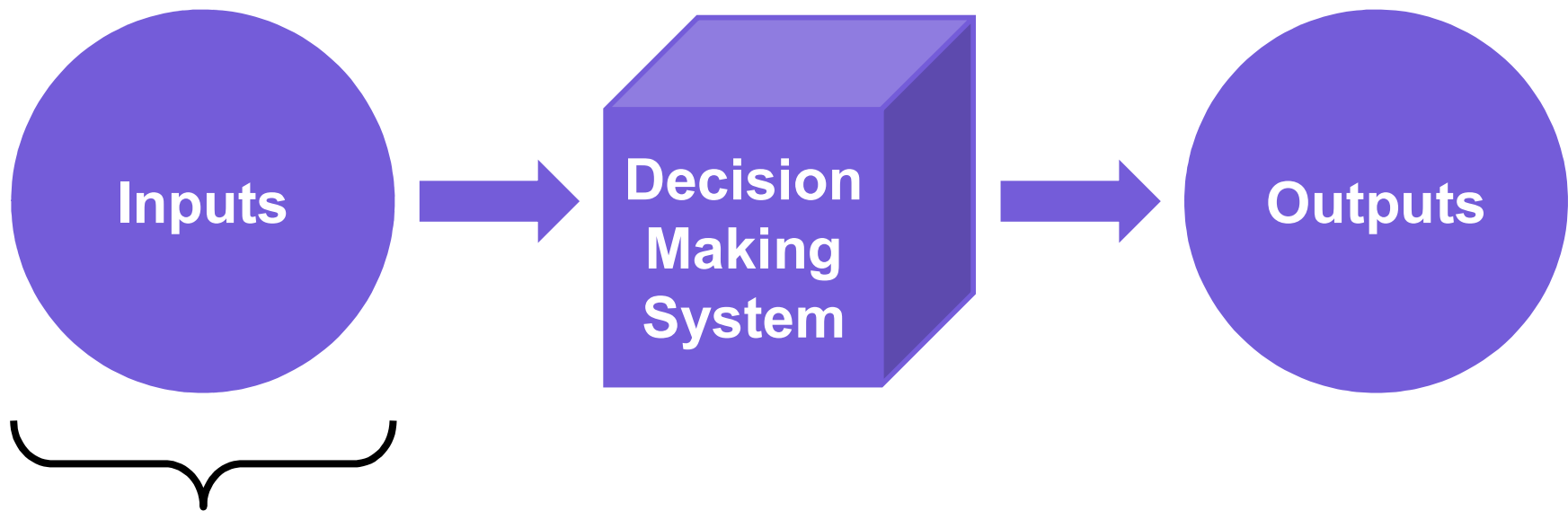
Algorithms help people make decisions about

- *Hiring*
- *Assigning social benefits*
- *Granting bail*

Are these algorithms **fair**?

Decision Making Pipeline

Example: **Granting bail**



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

This Talk

- Is it **fair** to use a feature?
- Why do people **perceive** some features as unfair?
- Do people **agree** in their fairness judgments?

This Talk

- Is it **fair** to use a feature?
- Why do people perceive some features as unfair?
- Do people agree in their fairness judgments?

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?

- **Parents' criminal history**
- **Education**
- ...

Case Study: COMPAS Tool

- **Helps judges** decide if a person should be granted **bail**

Input: Defendant's answers to the **COMPAS questionnaire**



Output: Prediction of the defendant's criminal risk

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

Human Judgments of Fairness

Rating the **fairness of using a feature**

Current Charges
Criminal History: self
Substance Abuse
Stability of Employment
Personality
Criminal Attitudes
Neighborhood Safety
Criminal History: others
Quality of Social Life
Education & School

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

This Talk

- Is it **fair** to use a feature?
- Why do people perceive features as unfair?
- Do people agree in their fairness judgments?

This Talk

- Is it fair to use a feature?
- **Why do people perceive features as unfair?**
- Do people agree in their fairness judgments?

Hypothesis I: Latent Properties of Features

Relevant?

Reliable?

Volitional?

Private?

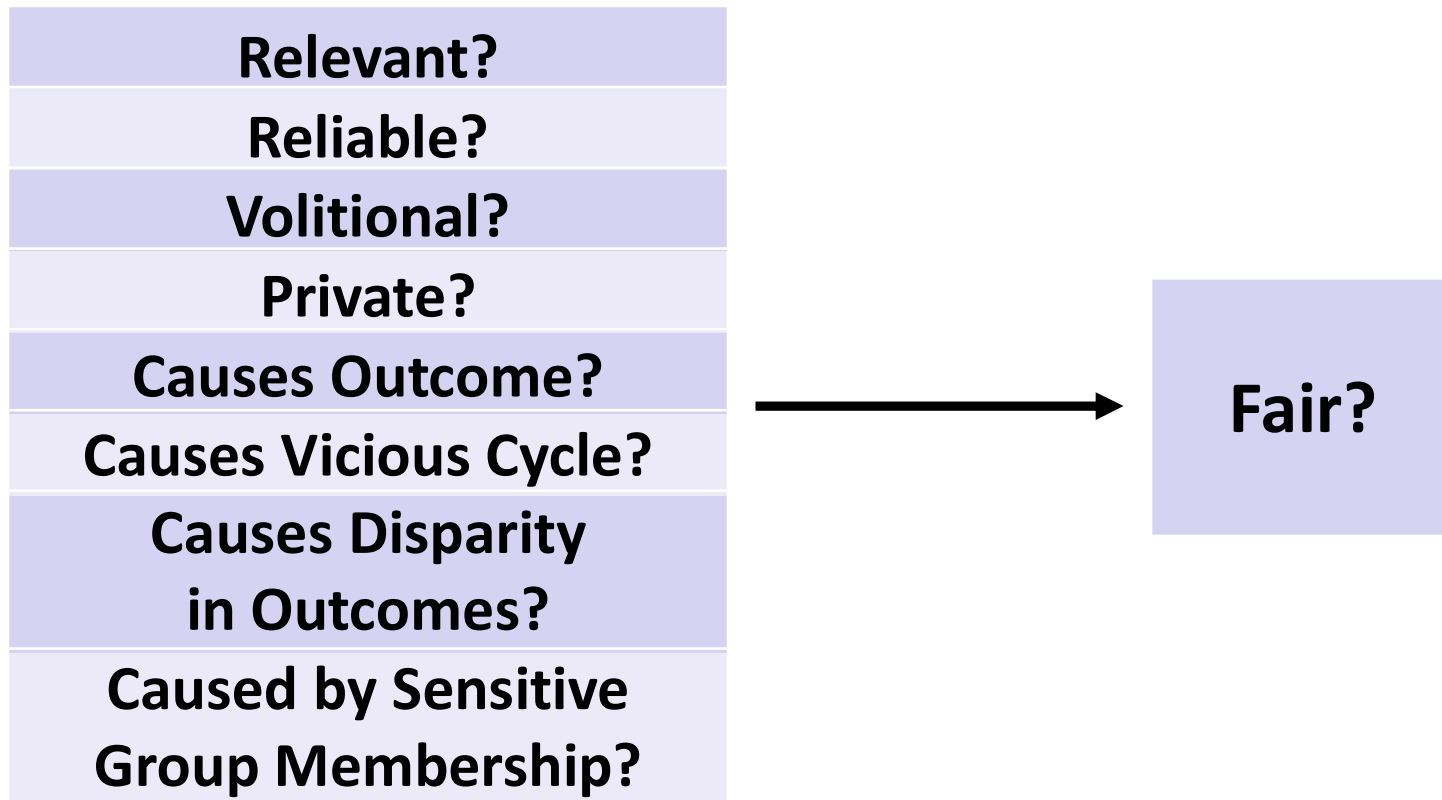
Causes Outcome?

Causes Vicious Cycle?

**Causes Disparity
in Outcomes?**

**Caused by Sensitive
Group Membership?**

Hypothesis II: From Latent Properties to Fairness



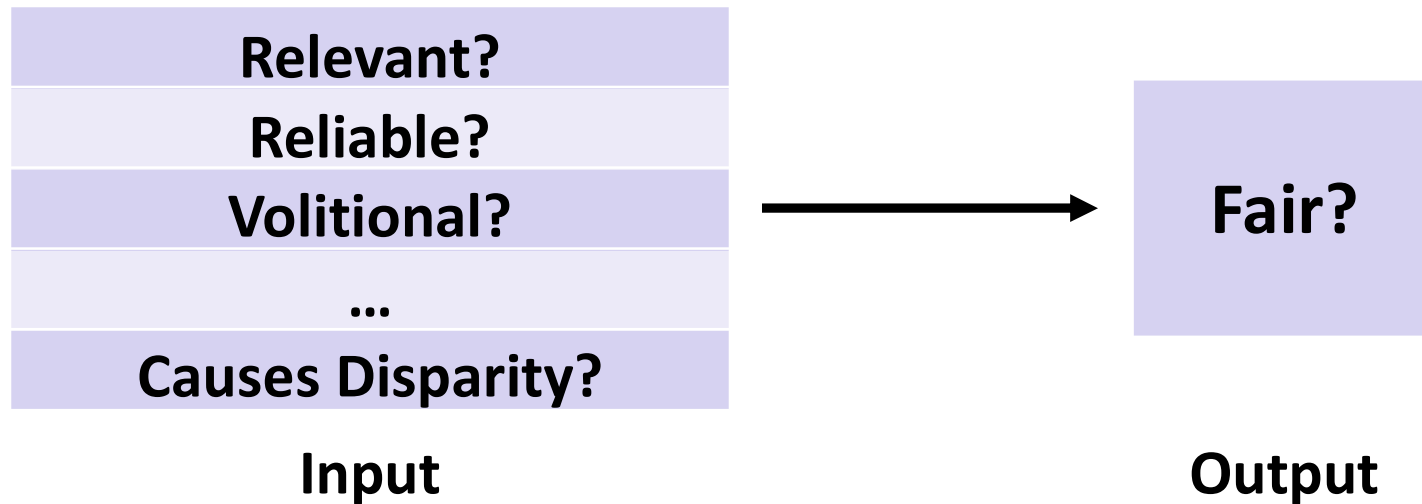
Reasons Behind Fairness Judgments

Why is it **(un)fair** to use a certain feature?

Relevance
Causes outcome
Reliability
Privacy
Volitionality
Vicious cycle
Causes disparity
Caused by sensitive
Other

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

Modeling Fairness Judgments



- We can predict fairness judgments with **88% accuracy**
- **Common fairness judgment heuristic** used by respondents
 - May depend on society: **interesting future work**

This Talk

- Is it fair to use a feature?
- **Why do people perceive features as unfair?**
 - **Fairness** of using a **feature** depends on **latent properties**
 - **Relevant? Volitional? Reliable? Private?**
- Do people agree in their fairness judgments?

This Talk

- Is it fair to use a feature?
- Why do people perceive features as unfair?
 - Fairness of using a feature depends on latent properties
 - Relevant? Volitional? Reliable? Private?
- Do people agree in their fairness judgments?

Disagreements in Fairness Judgments

Do people **agree** in their fairness judgments?

Current Charges

Criminal History: self

Substance Abuse

Stability of Employment

Personality

Criminal Attitudes

Neighborhood Safety

Criminal History: others

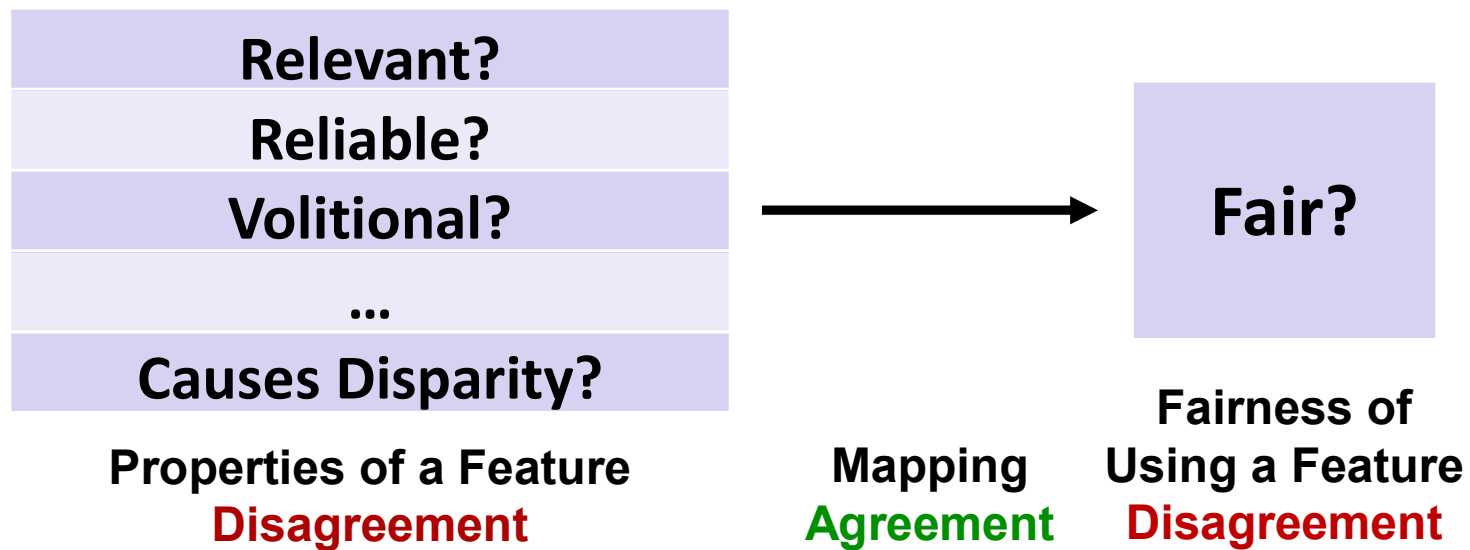
Quality of Social Life

Education & School

People often **disagree** in their fairness judgments

Causes of Disagreements in Fairness Judgments

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

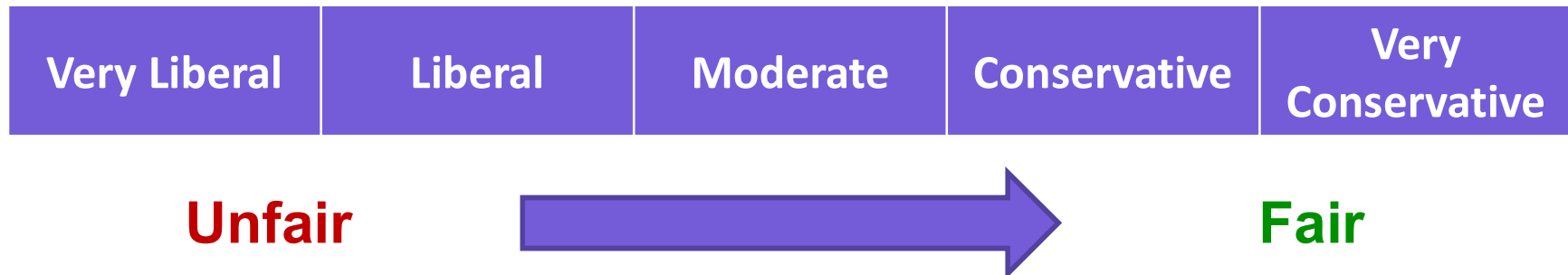


Disagreements in Latent Property Assessments?



Demographics and Fairness Judgments

- **Demographic factors:** *age, race, education, gender*
 - **No** statistically significant **differences**
- **Political views**
 - Statistically significant **differences**



- Many attributable to differences in **causal reasoning**

Summary

- The **fairness** of using a **feature** depends on its **latent properties**
- Fairness considerations go **beyond discrimination**
 - **Relevant? Volitional? Reliable? Private?**
- **Disagreement** in **fairness judgments**
 - **Agreement** in **mapping** from latent properties to fairness
 - **Disagreement** in **latent property** assessments
 - **Especially** those related to **causality**
 - Correlated with **ideological views** of people in the society

Future Directions

Cross-cultural studies of human perceptions of fairness

Disagreement in **latent property** assessments

- Can be **objectively** assessed?

Agreement in **mapping** from latent properties to fairness

- **Subjective** moral reasoning

Future directions for fair algorithmic decision making

- Moral reasoning about **mapping** gathered from **people**
- **Latent properties** assessed from **data**?

Additional Slides

Accounting For Fairness Judgments

Goal: Building decision making systems that account for **human perceptions of fairness (AAAI'18)**

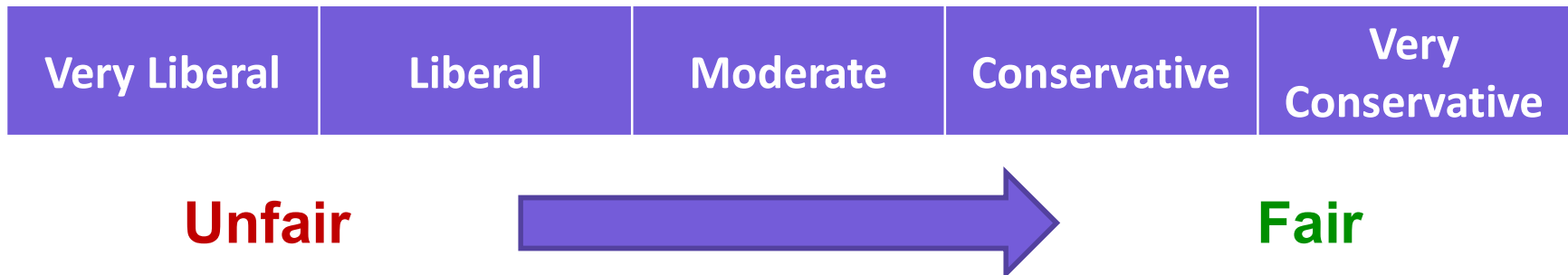
- Train **classifiers** which use **features** that
 - People **perceive as fair**
 - Achieve high **accuracy of prediction**

Demographics of People & Their Fairness Judgments

- So far, we examined the relationship between properties of **features** & fairness judgments
- Now, we consider the relationship between of **people** & fairness judgments
- **Properties of people** we consider
 - Demographic factors: *age, race, education, gender*
 - *Political leaning*

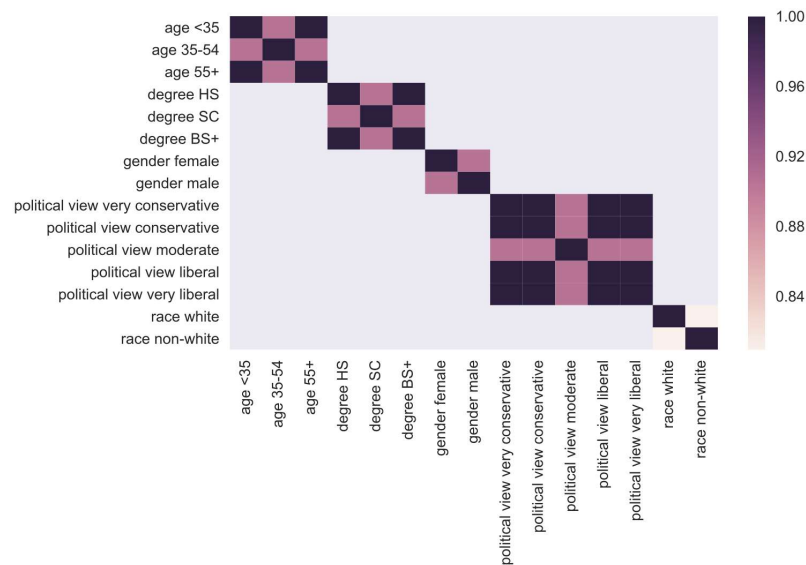
Differences in Fairness Judgments

- **Demographic factors:** *age, race, education, gender*
 - **No** statistically significant **differences**
- **Political views**
 - Statistically significant **differences**



Similarities in Fairness Judgments

- **Ranking** a set of **features** with respect to fairness
 - Derived from mean fairness ratings



Ranking consistent across different groups of people!

Properties of People & Their Fairness Judgments

- Deciding if a given feature fair to be used
 - **Disagreement** across people with different **political views**
 - More **right** leaning → more likely to consider a feature **fair**
- Ranking a set of features with respect to fairness
 - **Agreement** across **all** groups of people
- **General consensus** on parts of **algorithmic fairness**