Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction

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

[Part 2]
DATA.ML.381 Fairness in Big Data Management

[Team 4]

Mirva Pekkola

Niilo Pääkkönen

Nursat Sultana Kakon

1. Recap *(Motivation)*

The goal of this study is uncovering the moral reasoning behind people's perceptions towards fair decision making.

Complementary descriptive steps:

1. Pilot Survey 1: *Fairness Judgments & their Latent Reasons.*

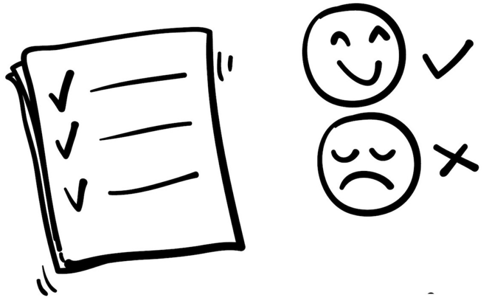
   Goal*: Learn whether respondents found the predictive features fair, and why they felt it was fair or unfair.*

2. Pilot Survey 2: *Latent Properties of Features*

   Goal*: Explore how people evaluate the latent properties of features from our framework. Here we asked no fairness-related questions, in order to control for the effect of asking about fairness on latent property ratings.*

3. Pre-test: Questionnaire Validity

   Goal: *Ensure meaningful interpretation of the questions by survey participants.*



1.1. Recap (*Main Survey Design*)

Main Survey: Fairness Judgments & Latent Properties of Features.

Goal*: Evaluate people's judgments about the latent properties of features were relevant to their judgments about the fairness of features.*

Example questions:

Q1: Please rate how much you agree with statements of the form <feature> has <inherent_property> .

Q2: How fair it is to use that <feature>?

Predictive Feature
Current Charges
Criminal History: self
Substance Abuse
Stability of Employment & Living Situation
Personality
Criminal Attitudes
Neighborhood Safety
Criminal History: family and friends
Quality of Social Life & Free Time
Education & School Behavior

Survey samples:

AMT (*Amazon Mechanical Turk*) **: 196**

SSI (*Survey Sampling International*) **: 380**

Total : 576

Latent Properties
Reliability
Relevance
Privacy
Volitionality
Causes Outcome
Causes Vicious Cycle
Causes Disparity in Outcomes
Caused by Sensitive Group Membership

2. Fairness Judgements Analysis

First

Across features: Compare respondents’ judgments on the fairness of using different *features* in our algorithmic decision-making scenario.

Second

Across users: Explore the degree to which they reach consensus in their judgments on the usage of any given *feature*.

2.1. Fairness Judgements *across features*

Feature	1	2	3	4	5	6	7
Current Charges	0,01	0,01	0,01	0,03	0,12	0,18	0,65
Criminal History: self	0,02	0,01	0,01	0,03	0,08	0,22	0,64
Substance Abuse	0,08	0,07	0,10	0,07	0,26	0,22	0,20
Stability of Employment	0,13	0,05	0,11	0,09	0,26	0,24	0,12
Personality	0,16	0,18	0,11	0,10	0,22	0,12	0,12
Criminal Attitudes	0,22	0,12	0,16	0,09	0,20	0,11	0,09
Neighborhood Safety	0,28	0,21	0,15	0,07	0,12	0,10	0,08
Criminal History: family and friends	0,38	0,21	0,09	0,07	0,13	0,10	0,03
Quality of Social Life & Free Time	0,38	0,20	0,12	0,07	0,12	0,08	0,03
Education & School Behavior	0,34	0,22	0,14	0,08	0,13	0,06	0,03

Figure: Fraction of Respondent's Rating on Features (AMT respondents)

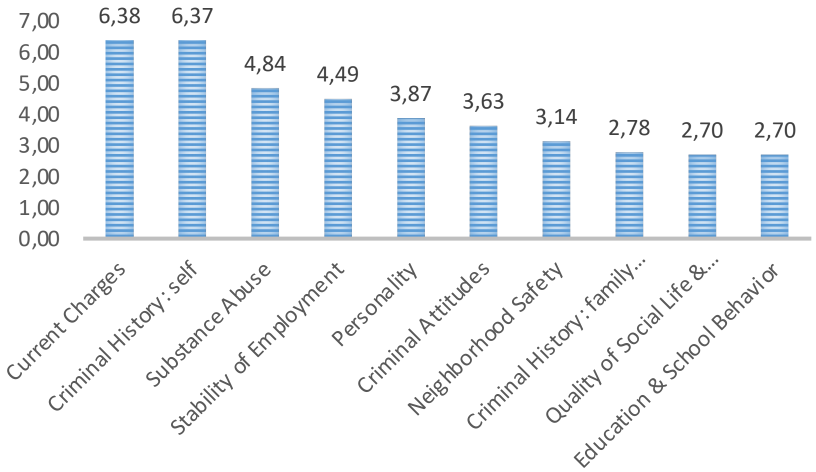
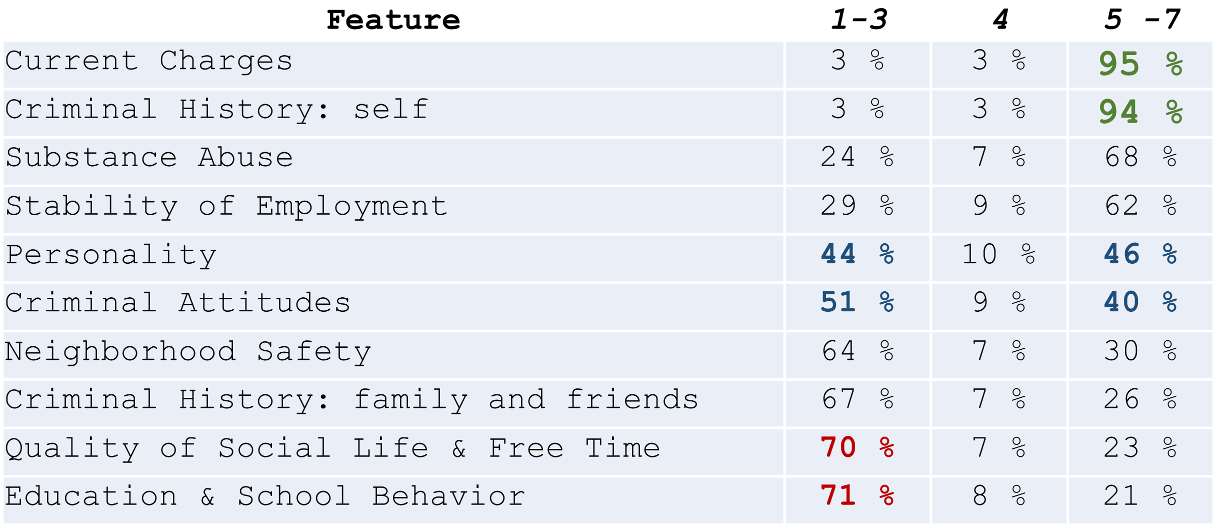


Figure: Mean of Respondent's Rating Features (AMT respondents)

- *'Current Charges'* & *'Criminal History: self'* is mostly fair to use.
- *'Education & School Behavior'*, *'Quality of Social Life & Free Time'*, & *'Criminal History: Family and Friends'*, are rated as somewhat unfair to use.
- More than half of the respondents judged five out of ten *features* as unfair (rating 1-3).

None of the *features* directly capture sensitive group membership information. The use of these *features* are not restricted by anti-discrimination laws. So, we can say that algorithmic unfairness concerns goes beyond discrimination.

2.2. Fairness Judgements *across users*



Feature	***1-3***	***4***	***5 -7***
Current Charges	3 %	3 %	**95 %**
Criminal History: self	3 %	3 %	**94 %**
Substance Abuse	24 %	7 %	68 %
Stability of Employment	29 %	9 %	62 %
Personality	**44 %**	10 %	**46 %**
Criminal Attitudes	**51 %**	9 %	**40 %**
Neighborhood Safety	64 %	7 %	30 %
Criminal History: family and friends	67 %	7 %	26 %
Quality of Social Life & Free Time	**70 %**	7 %	23 %
Education & School Behavior	**71 %**	8 %	21 %

- Consensus between respondents achieved high on *two (‘Current Charges’, ‘Criminal History: self’)* out of the ten *features*.
- *‘Personality’* and ‘*Criminal Attitudes*’, are very low consensus, with neither fair nor unfair.
- Many of the remaining *features*, there is a reasonable consensus among respondents. Two-thirds majority considering the feature either fair (5-7) or unfair (1-3).

2.2.1. Shannon Entropy

Goal:

Measure the consensus amongst people's ratings of fairness & *latent property* values which calculated over the probability distributions on the answers from the respondents.

Measurement:

1 – NSE (Normalized Shannon Entropy) between 0 as *complete disagreement* and 1 as *complete consensus*.

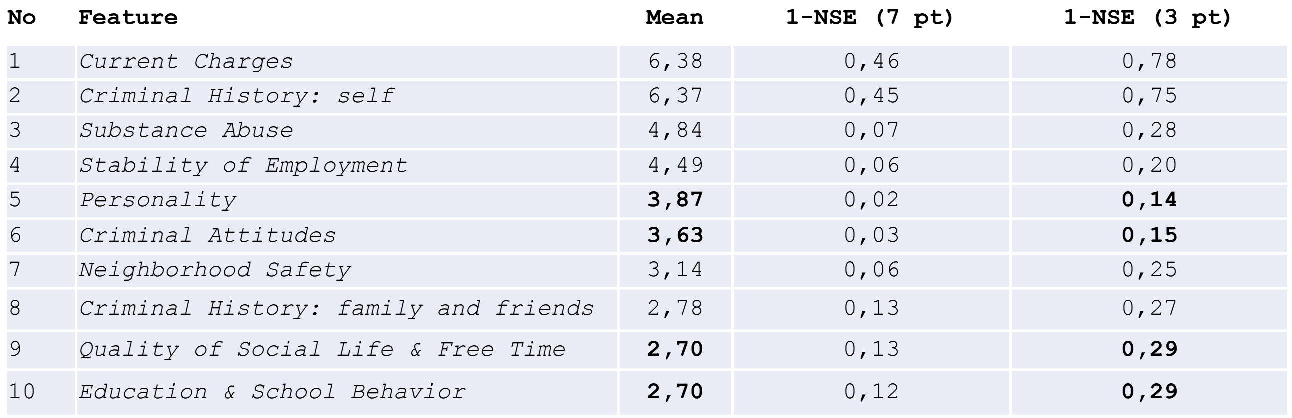
$$H(X) = -\sum_{i=1}^{n} P(x_i) log_2 P(x_i)$$

$$H_n(X) = -\sum_{i=1}^{n} \frac{P(x_i) log_2 P(x_i)}{log_2 n}$$

2.2.1. Shannon Entropy (*example*)

			Shanon Entrophy H(X)	**Normalized Shanon Entropy $\mathbf{H_n(X)}$**
AAAAAAAA	Low	$P(A) = \frac{8}{8} = 1$	$-1\log_2(1)$ $= 0$	0
BAADABAC	Medium	$P(A) = \frac{4}{8} = \frac{1}{2}$ $P(B) = \frac{2}{8} = \frac{1}{4}$ $P(C) = \frac{1}{8}$ $P(D) = \frac{1}{8}$	$-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{8}\log_2\left(\frac{1}{8}\right) - \frac{1}{8}\log_2\left(\frac{1}{8}\right)$ $= 1.75$	$\frac{1.75}{2} = 0.875$
DBCADACB	High	$P(A) = \frac{2}{8} = \frac{1}{4}$ $P(B) = \frac{2}{8} = \frac{1}{4}$ $P(C) = \frac{2}{8} = \frac{1}{4}$ $P(D) = \frac{2}{8} = \frac{1}{4}$	$-\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$ $= 2$	$\frac{2}{2} = 1$

2.2. Fairness Judgements *across users (cont.)*



No	Feature	Mean	1-NSE (7 pt)	1-NSE (3 pt)
1	*Current Charges*	6,38	0,46	0,78
2	*Criminal History: self*	6,37	0,45	0,75
3	*Substance Abuse*	4,84	0,07	0,28
4	*Stability of Employment*	4,49	0,06	0,20
5	*Personality*	**3,87**	0,02	**0,14**
6	*Criminal Attitudes*	**3,63**	0,03	**0,15**
7	*Neighborhood Safety*	3,14	0,06	0,25
8	*Criminal History: family and friends*	2,78	0,13	0,27
9	*Quality of Social Life & Free Time*	**2,70**	0,13	**0,29**
10	*Education & School Behavior*	**2,70**	0,12	**0,29**

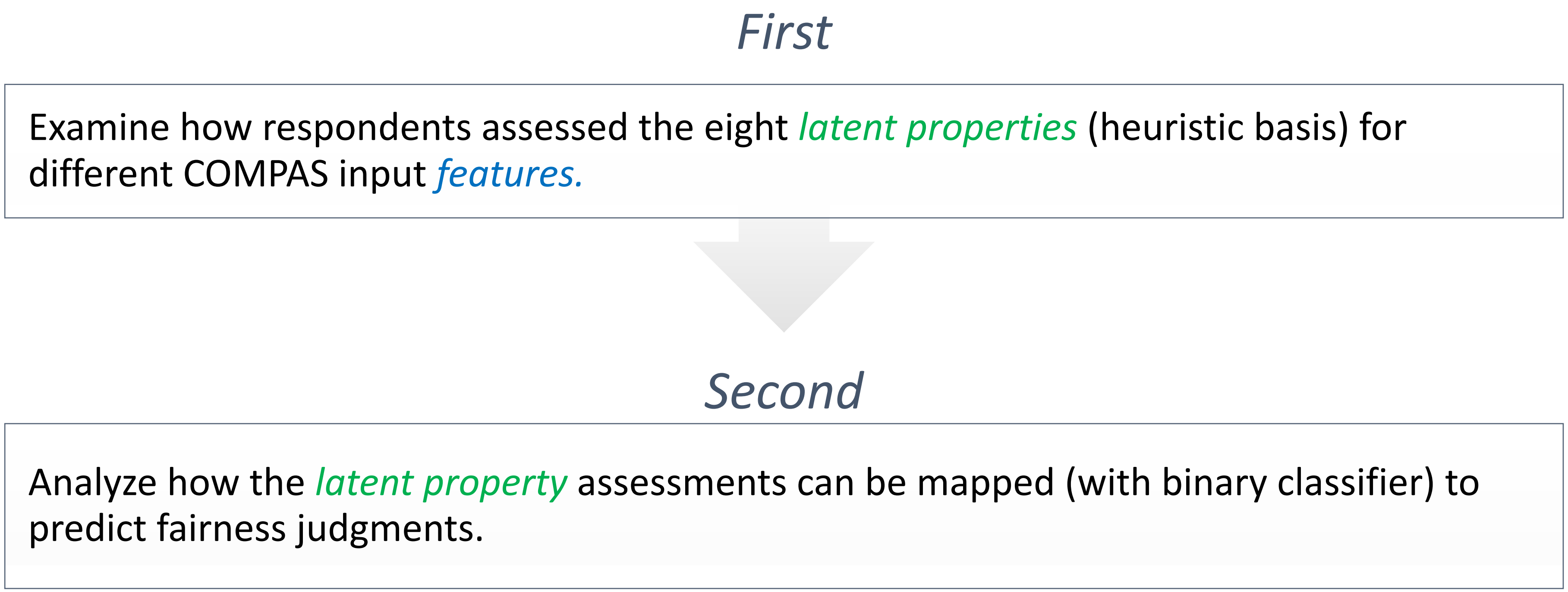
Figure: Consensus in fairness judgments for different features measured using 1 - NSE.

- *Features* with mean ratings close to neutral (rating 4) such as '*Personality*' & '*Criminal Attitudes*' exhibit little consensus, with judgments of respondents spread across the entire rating spectrum from 1 through 7.
- Low consensus on *features* rated as least fair to use, such as '*Education* & *School Behavior*'.

It is possible that societal consensus for or against the use of these *predictive features* is still evolving, unlike the broad consensus against using features (e.g: race or gender) that has been codified in anti-discrimination laws.

3. Fairness Reasoning Analysis

Goal: Explore the possible causes of lack of consensus in respondents' fairness judgments.



First

Examine how respondents assessed the eight *latent properties* (heuristic basis) for different COMPAS input *features.*

Second

Analyze how the *latent property* assessments can be mapped (with binary classifier) to predict fairness judgments.

3.1. Latent Property Assessments

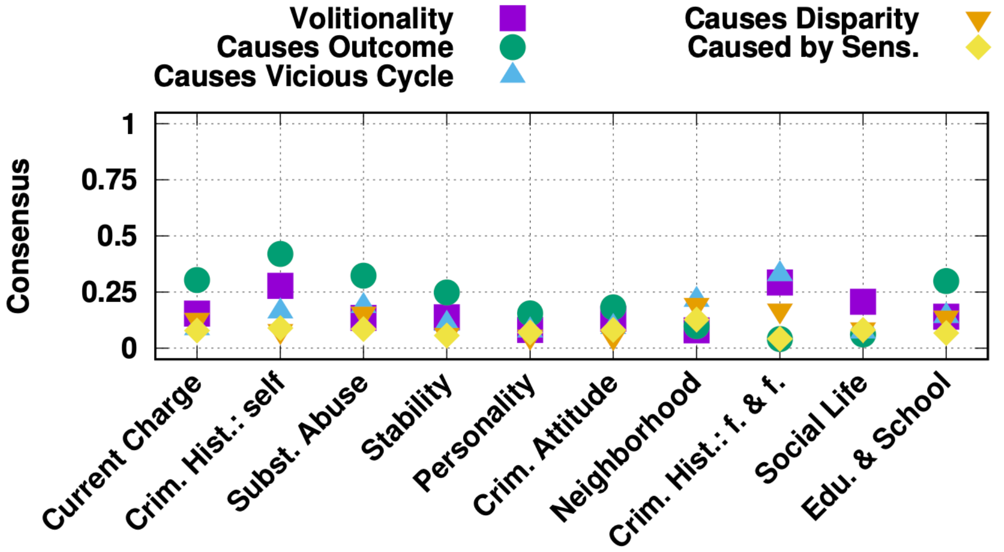


Figure: Comparatively higher degree of consensus in fairness judgments & assessments of latent properties.

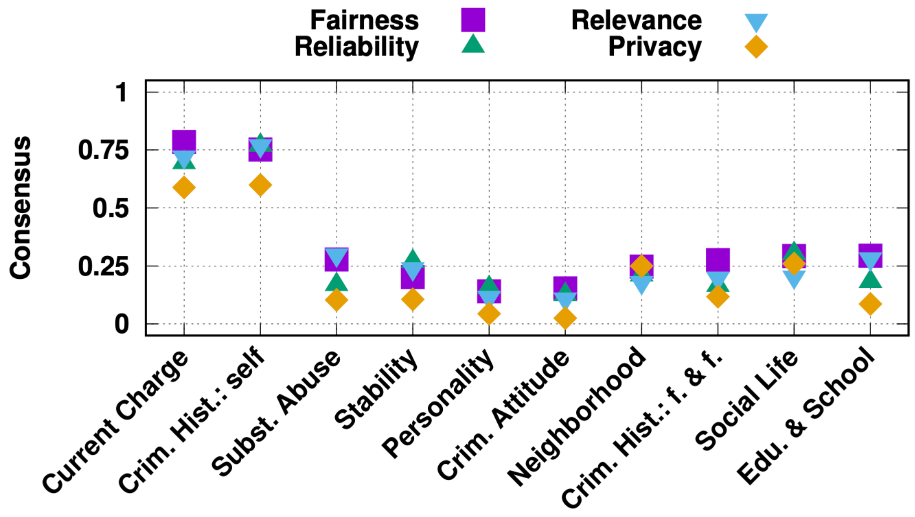


Figure: Comparatively lower degree of consensus in fairness judgments & assessments of latent properties.

Causal properties:

- Generally, people tend to disagree in their assessments of all *latent properties* for at least one or more *features*.
- The *properties* related to causality appear particularly controversial & has low consensus (< 0.5) for all input *features*.
- Consensus around *latent properties* related to any *feature's* potential to cause discrimination is lowest (< 0.2).

Non-Causal properties:

- *Latent properties* based on non-causality achieve high consensus (> 0.5) on at least some input *features*.
- High consensus in these *latent properties* estimate high consensus in respondents' fairness judgments over the corresponding COMPAS input *features*.

3.2 Predicting Fairness Judgments

Goal: Quantify the predictive power of the latent property assessments on fairness judgments.

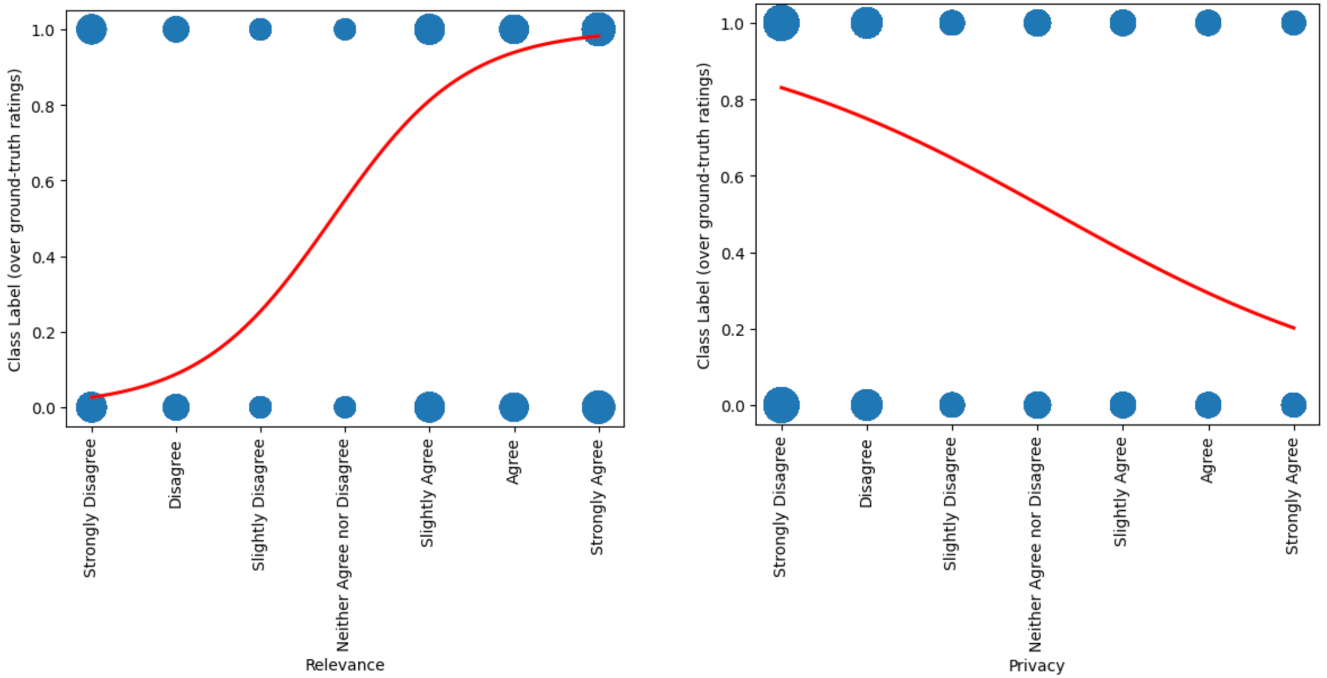
Model fairness reasoning of our respondents to predict the fairness judgements about an input feature based on the feature's latent properties.

Binary Classifier

Predict if a feature will be considered fair (*completely, mostly, slightly, neutral*) or unfair (*completely, mostly, slightly*), based on the values of its latent properties.

Training data: respondents' evaluations of latent properties & binarized evaluations of fairness.

Analysis: Train a logistic regression model with L2 regularization.



Ratings	relevance	privacy
1	331	**463**
2	242	345
3	173	221
4	163	253
5	334	238
6	308	231
7	**409**	209
Total	*1960*	*1960*

Figure: Linear regression model with 'Relevance'& 'Privacy' properties across all the features.

3.2.1. Logistic Model Accuracy

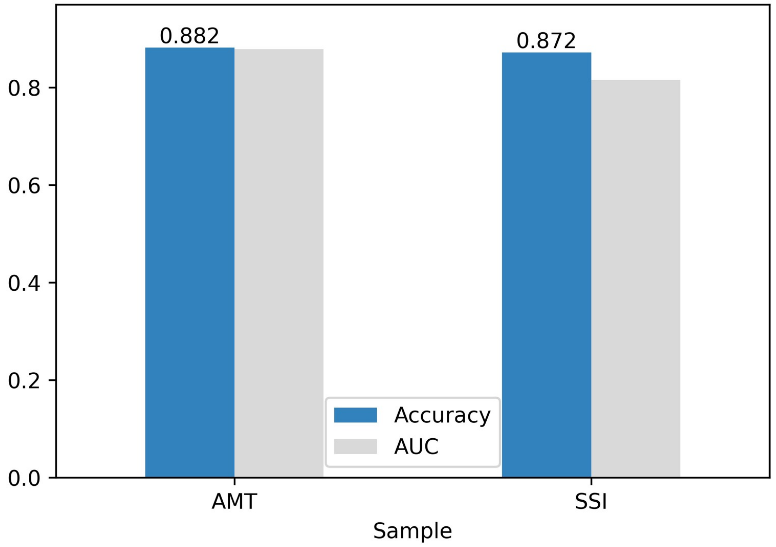


Figure: Accuracy & AUC of binary classifier predicting feature (including neutral fairness judgments)

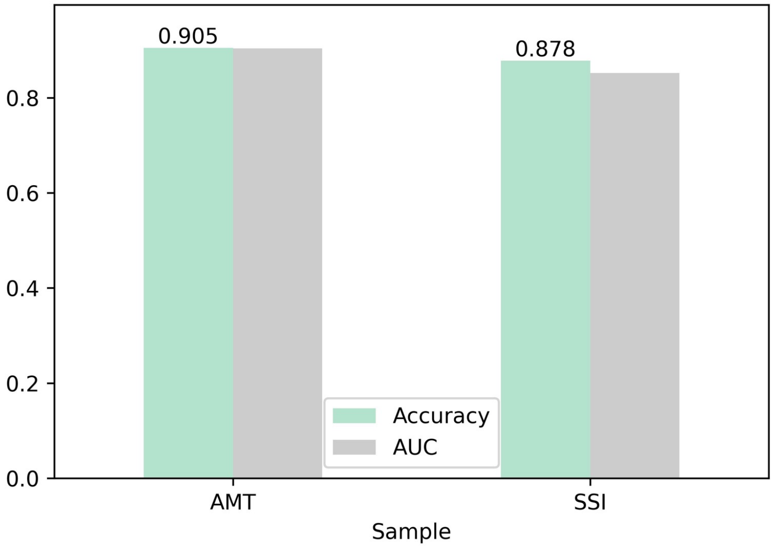
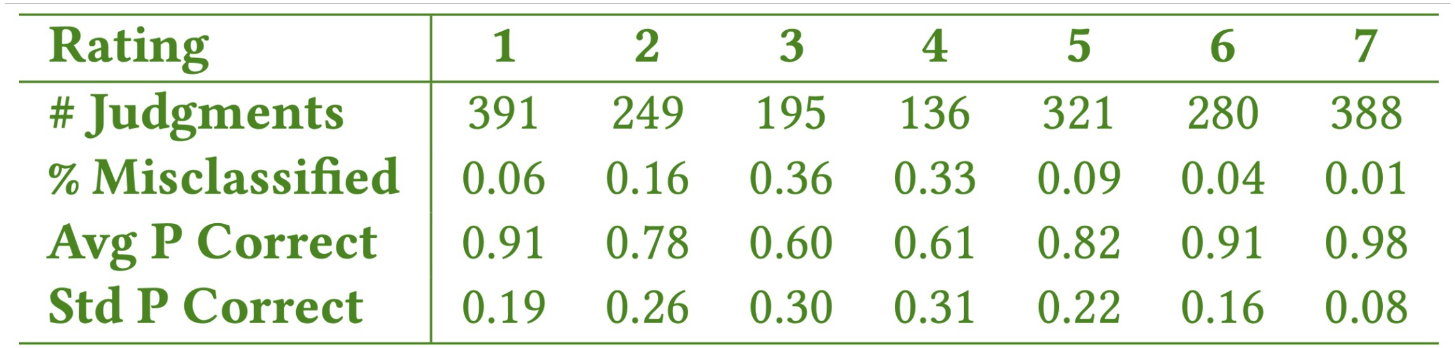


Figure: Accuracy & AUC of binary classifier predicting feature (excluding neutral fairness judgments)

- Classifier can make highly accurate (>= 80% accuracy) for most (> 85%) of all AMT respondents.
- The classifier achieves very high accuracy:
  - 88.2% with AUC 0.879 for AMT
  - 87.2% with AUC 0.816 for SSI

  when predicting respondents’ fairness judgments based on the underlying *property* ratings they assigned.

3.2.2. Logistic Model Evaluation



Rating	1	2	3	4	5	6	7
# Judgments	391	249	195	136	321	280	388
% Misclassified	0.06	0.16	0.36	0.33	0.09	0.04	0.01
Avg P Correct	0.91	0.78	0.60	0.61	0.82	0.91	0.98
Std P Correct	0.19	0.26	0.30	0.31	0.22	0.16	0.08

Row1: Total number of entries that received a certain rating

Row2: Fraction of ratings that misclassified by model

Row3: Avg. probability value that assign to correct class

Row4: Std. probability value that assign to correct class

Figure: Accuracy & AUC of binary classifier predicting feature (including neutral fairness judgments)

By studying how respondents' judgments are distributed over the ground-truth fairness ratings (AMT respondents), there are a few misclassifications of the model:

- The model reaches high accuracy on features that were rated very unfair (rating: 1-2) or very fair (rating: 6-7).
- On the neutral ratings (rating: 4) performance is close to random.

3.3. Fairness Reasoning Findings

We can say that the proposed framework is effective at modeling people's moral judgements about fairness.

The high accuracy of predictions for most respondents strongly suggests that:

i. Eight *latent properties* are sufficient to explain user’s fairness judgements.

ii. Most respondents are using similar reasoning at reaching their fairness judgements even if they would disagree in the assesment of *latent properties* at the beginning.

3.4. Relative Impact of Latent Properties

Goal: Analyze odds ratios (log-adjusted regression coefficient) to estimate the relative importance of our latent properties on respondents' fairness judgments.

- ‘*Relevance*’ had the strongest effect, with the odds of respondents rating the scenario as fair increasing 2.47 times for every point higher they rated it on a 7-point Likert scale.
- Respondents were more likely to judge the use of a *feature* fair, the more they felt that the *feature* is ‘*reliable*’, ‘*relevant*’, ‘*volitional* ’, & ‘*caused the outcome*’.
- Respondents were less likely to judge the use of a *feature* as fair if they felt that the *feature* is ‘*privacy*’ sensitive, or resulted in a ‘*vicious cycle*’.

Upon comparison, it is found that the SSI respondents are more likely to rate a scenario as fair than AMT respondents.

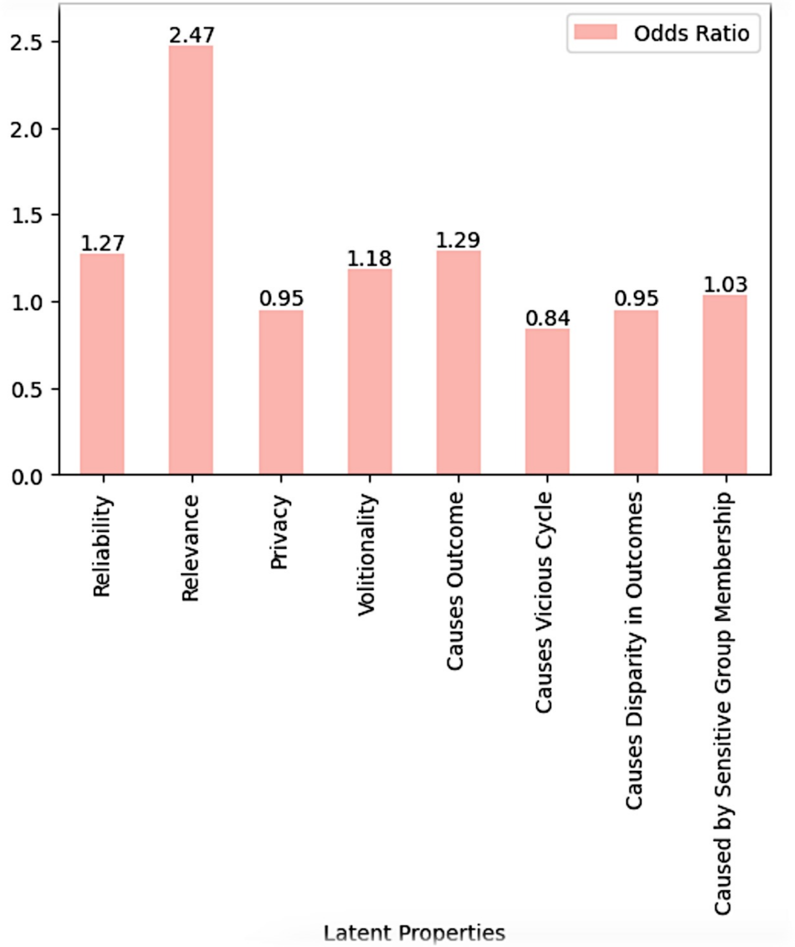


Figure: Odds ratios for the logistic regression model.

4. Fairness Disagreements Analysis

Comparison of the consensus in the predicted fairness judgments *with* the consensus in the ground truth fairness judgments, assigned by respondents.

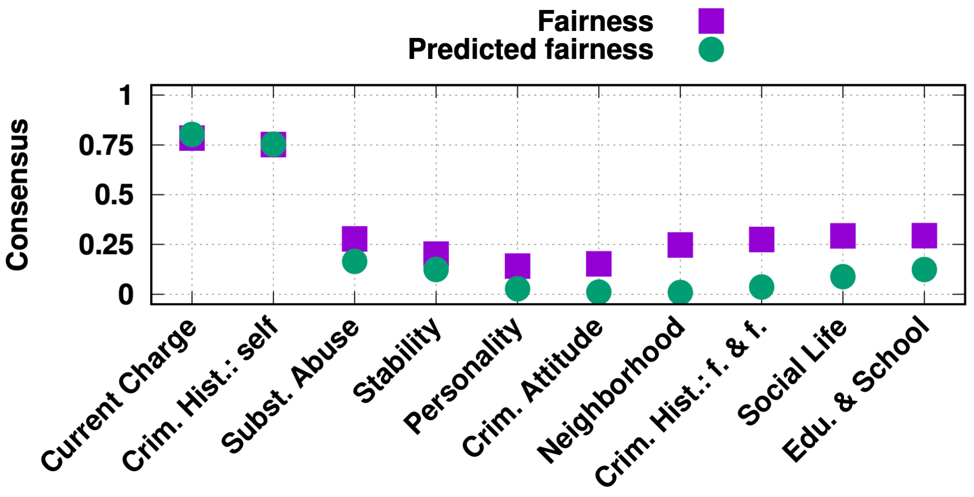


Figure: Comparison between consensus in fairness predictions by model with consensus in ground truth fairness judgments

Findings:

- Fairness disagreements likely arise out of disagreements in how people assess *latent properties* of *features* rather than how they use the *latent properties* to reason about fairness of using the *features*.
- Significant part of the lack of consensus in human fairness judgments to the differences in people's assessments of *latent properties.*

5. Conclusion

I. People’s concerns about the *unfairness* of using a *feature* extend beyond discrimination, including consideration of *latent properties* (e.g. ‘relevance’, ‘reliability’) to the decision-making scenario.

II. There are considerable disagreements on which *features* different people perceive as *unfair* to use.

III. Lack of consensus can be attributed to disagreements in how people assess the *latent properties* of the *features*, particularly those related to causal relationships between input *features* and their causal influence on outcomes.

IV. Different people appear to share a common heuristic when reaching their *fairness* judgment from their assessments of *latent properties*.

Thank You