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

[Part 1]

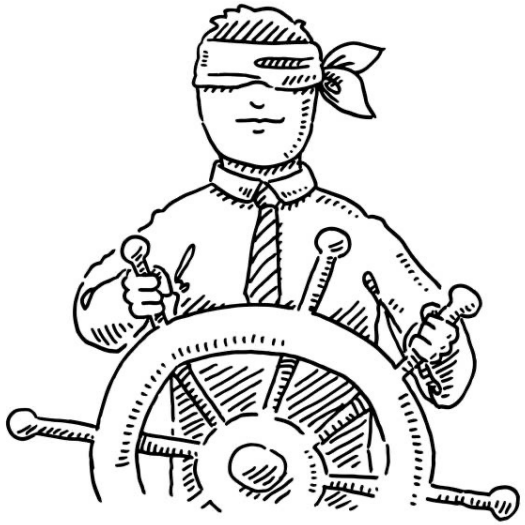
DATA.ML.381 Fairness in Big Data Management

[**Team 4**]
Mirva Pekkola
Niilo Pääkkönen
Nursat Sultana Kakon

# Background

Existing studies on algorithmic fairness normatively prescribe how fair decisions ought to be made.

Most *prior work* begin by defining how fair decisions should be made, assuming that there is societal consensus around what constitutes fair decision making.

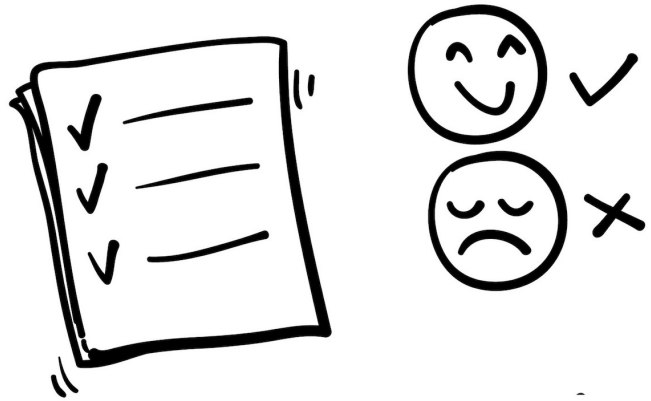


# Motivation

This study conducted by a complementary descriptive approach (surveys) towards fair decision making.

> This study proposes a framework that helps to understand *why people perceive certain features as fair or unfair to be used in algorithms*.

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



# Judging Features in Fairness

COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*):

an automated decision-support software package that integrates risk and needs assessment with several other domains, including sentencing decisions, treatment and case management, and recidivism outcomes.

It is developed and owned by Northpointe used by U.S. courts to assess the likelihood of a defendant becoming a recidivist.


This study assess the fairness of using the different input features to COMPAS for making bail decisions.

# *Sample* Parole Document of COMPAS

## Risk Assessment

| PERSON | | | |
|---|---|---|---|
| Name: [redacted] | Offender #: [redacted] | | DOB: [redacted] |
| [redacted] | Gender: Male | Marital Status: Single | Agency: DAI |

| ASSESSMENT INFORMATION | | | |
|---|---|---|---|
| Case Identifier: [redacted] | Scale Set: Wisconsin Core - Community Language | Screener: [redacted] | Screening Date: [redacted] |

### Current Charges

| | | | |
|---|---|---|---|
| ☐ Homicide | ☑ Weapons | ☑ Assault | ☐ Arson |
| ☐ Robbery | ☐ Burglary | ☐ Property/Larceny | ☐ Fraud |
| ☐ Drug Trafficking/Sales | ☐ Drug Possession/Use | ☐ DUI/OUIL | ☑ Other |
| ☐ Sex Offense with Force | ☐ Sex Offense w/o Force | | |

1. Do any current offenses involve family violence?
   ☑ No ☐ Yes
2. Which offense category represents the most serious current offense?
   ☐ Misdemeanor ☐ Non-violent Felony ☑ Violent Felony
3. Was this person on probation or parole at the time of the current offense?
   ☑ Probation ☐ Parole ☐ Both ☐ Neither
4. Based on the screener's observations, is this person a suspected or admitted gang member?
   ☐ No ☑ Yes
5. Number of pending charges or holds?
   ☑ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4+
6. Is the current top charge felony property or fraud?
   ☑ No ☐ Yes

### Criminal History

**Exclude the current case for these questions.**

# Predictive Feature in Fairness assessment

| Predictive Feature | Example Question |
|---|---|
| Current Charges | Are you currently charged with a misdemeanor, non-violent felony or violent felony? |
| Criminal History: self | How many times have you violated your parole? |
| Substance Abuse | Did you use heroin, cocaine, crack or meth as a juvenile? |
| Stability of Employment & Living Situation | How often do you have trouble paying bills? |
| Personality | Do you have the ability to “sweet talk” people into getting what you want? |
| Criminal Attitudes | Do you think that a hungry person has a right to steal? |
| Neighborhood Safety | Is there much crime in your neighborhood? |
| Criminal History: family and friends | How many of your friends have ever been arrested? |
| Quality of Social Life & Free Time | Do you often feel left out of things? |
| Education & School Behavior | What were your usual grades in high school? |

# Latent Properties in Fairness assessment

| Latent Properties | Example Question |
|---|---|
| Reliability | “Do you think that a hungry person has a right to steal?” |
| Relevance | “What were your usual grades in high school?” |
| Privacy | “Did you use heroin, cocaine, crack, or meth as a juvenile?” |
| Volitionality | “Was your father or mother ever arrested?” |
| Causes Outcome | “Are you currently charged with a misdemeanor, non-violent felony or violent felony?” |
| Causes Vicious Cycle | “Does any of your friends have ever been arrested?” |
| Causes Disparity in Outcomes | “Is there much crime in your neighborhood?” |
| Caused by Sensitive Group Membership | “How many of your friends have ever been arrested?” |

# *Pilot Survey 1*: Survey Design

Pilot Survey 1: *Fairness Judgments and Their Latent Reasons.*

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

*Example questions:*

Q1: Please rate how much you agree with the following statement: It is fair to determine if a person can be released on bail using information about their <feature>.

7 point Likert scale, where 1 = Strongly disagree, 7 = Strongly agree.

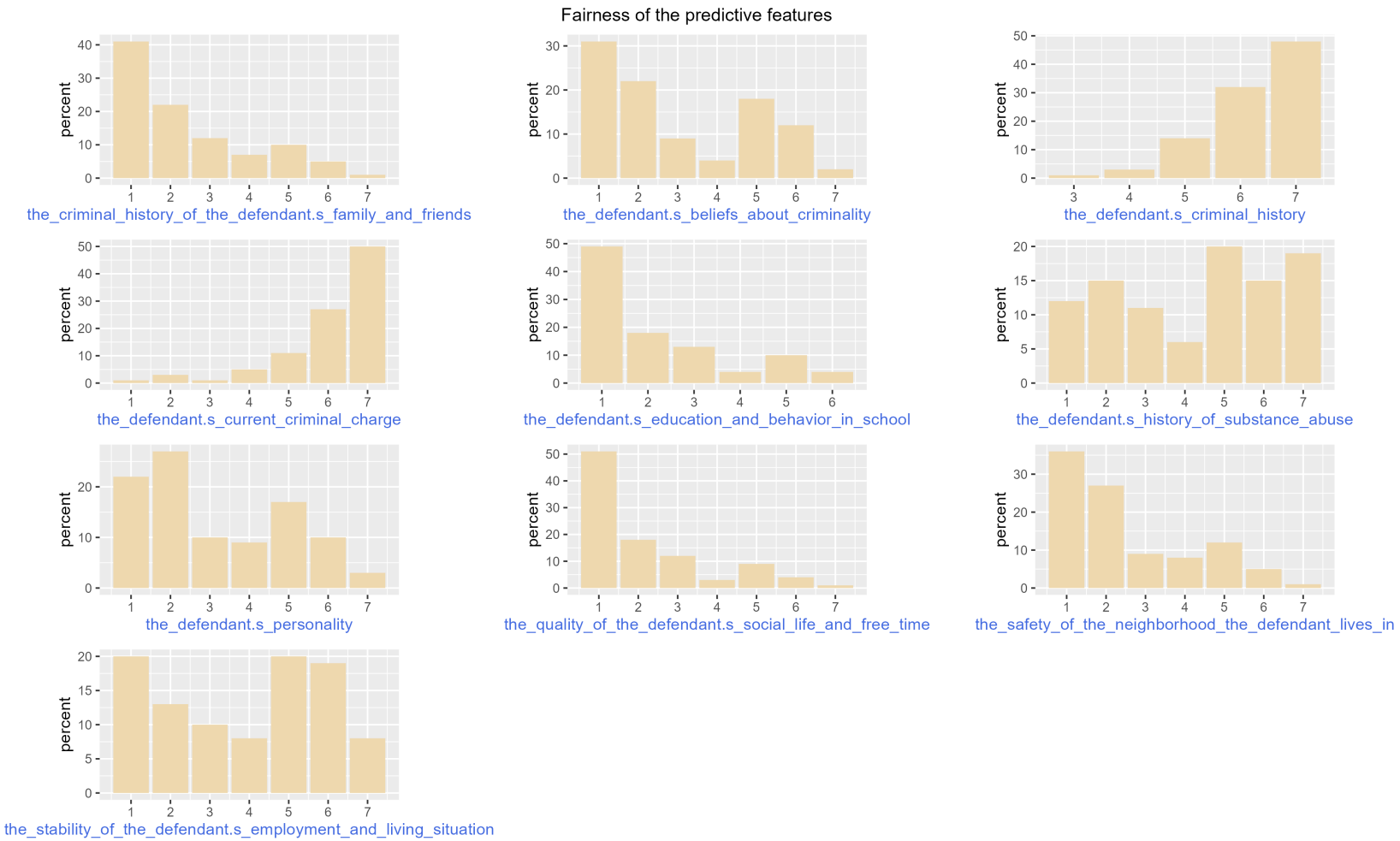
Q2: Select the reason(s) [<inherent_property> + "other"] for why <feature> was fair or unfair.

If their answer was in [1,4], we show them questions asking them why they thought it was unfair.
If their answer was in [4,7], we show them questions asking them why they thought it was fair.
if their answer was = 4, we show them both sets of questions.

| 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 |

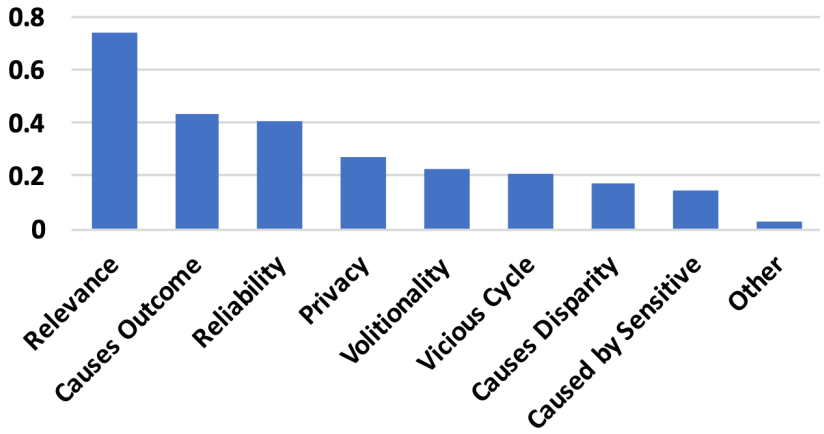
| Latent Properties |
|---|
| Reliability |
| Relevance |
| Privacy |
| Volitionality |
| Causes Outcome |
| Causes Vicious Cycle |
| Causes Disparity in Outcomes |
| Caused by Sensitive Group Membership |

# *Pilot Survey 1*: Takeaways



# *Pilot Survey 1*: Takeaways

- Each property was used by at least 15% of respondents (*among 98 participants*) to explain why they rated a particular feature as fair or unfair.
  - '*Relevance*' used in 74% of the responses.
  - '*Causes Outcome*' & '*Reliability*' used in more than 40% of the responses.
- Less than 3% of respondents selected the "*Other*" option for why they judged a scenario as fair or unfair.
  - Thematic analysis of responses reveals that the majority of these responses from the "*Other*" option still map to one of our eight properties.



# *Pilot Survey 2*: Survey Design

Pilot Survey 2: *Latent Properties of Features*

Goal*: to 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.*

*Example questions:*

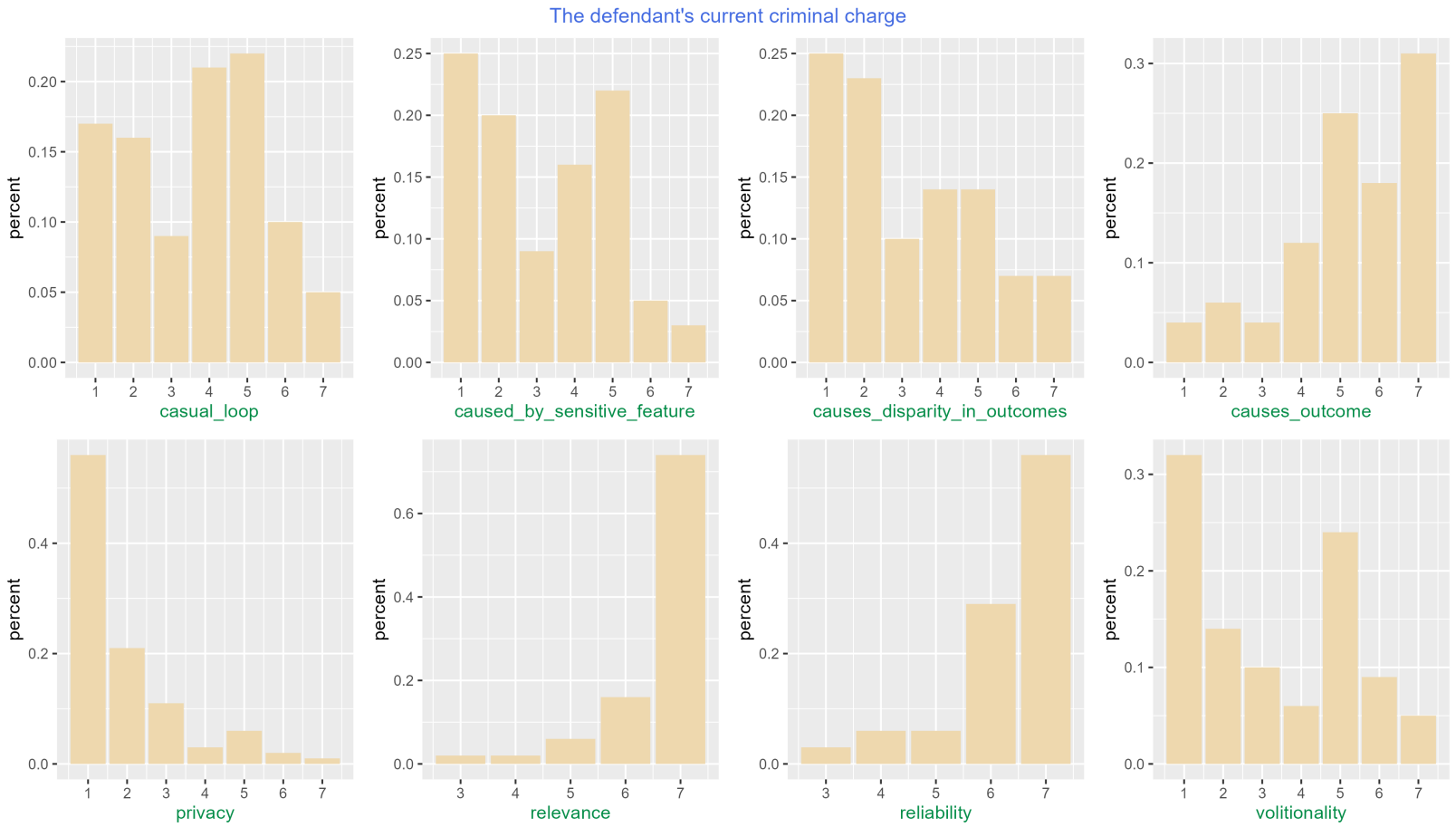
Q1: Please rate how much you agree with statements of the form <feature> has <inherent_property> .

7-point Likert scale, where 1 = Strongly disagree, 7 = Strongly agree.

| 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 |

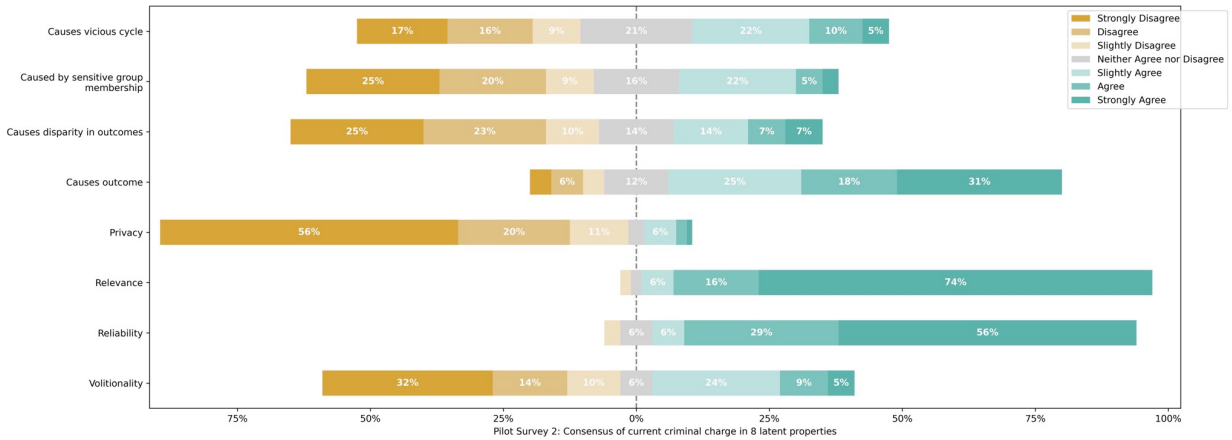
| Latent Properties |
|---|
| Reliability |
| Relevance |
| Privacy |
| Volitionality |
| Causes Outcome |
| Causes Vicious Cycle |
| Causes Disparity in Outcomes |
| Caused by Sensitive Group Membership |

# *Pilot Survey 2*: Takeaways



# *Pilot Survey 2*: Takeaways

A comparative analysis has been conducted between the results of (i) Pilot survey 2 (*100 respondents*), & (ii) the Main survey (*576 respondents*) to quantify this bias with KL-divergence.




- The KL-divergence from (i) to (ii) is very low
  - below 0.1 for 90% of the questions
  - below 0.14 for the remaining 10%

- Pilot survey 2 was used to examine latent property assessments without of fairness factor.
- Overall, the assessments of latent property values are minimally affected by questions about fairness.

# *Pre-test*: Questionnaire Validity

Goal: Ensuring meaningful interpretation of the questions by survey participants.

Pre-tested the main survey questionnaires using cognitive interviews which involve asking participants to think aloud to the researcher as they take the survey.

Participants: five demographically diverse respondents, recruited by *Prolific, a* crowdsourcing platform.

# *Pre-test*: Takeaways

- Based on participant feedback, iteratively refined the questionnaires to ensured validity.
- Randomized the order of questions described to mitigate the effects of order bias.
  - The order in which the features and latent properties were presented were randomized.
  - The order whether the fairness question was presented before or after the questions about the latent properties.
- Included an attention-check question to ensure that participants were thoughtfully answering the questionnaire items.
- Collected final sample for analysis.

# *Main Survey:* Survey Design

Main Survey: Fairness Judgments and Latent Properties of Features.

Goal*: To evaluate whether people's judgments about the latent properties of features proposed in our framework 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> .

7 point Likert scale, where 1 = Strongly disagree, 7 = Strongly agree.

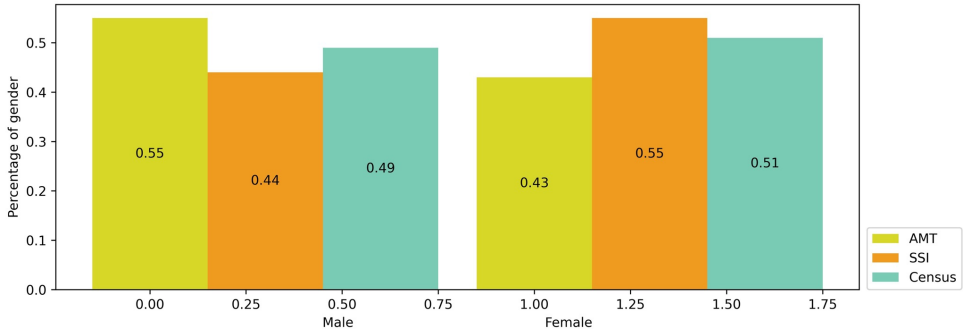
Q2: How fair it is to use that <feature>?

Randomly either right before or right after the questions about inherent properties.

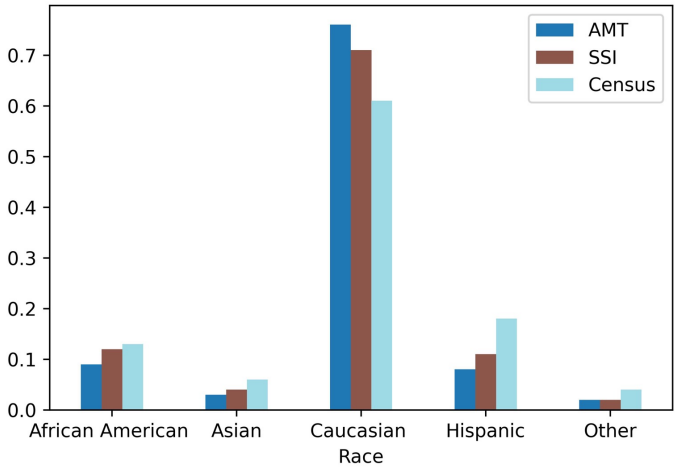
| Predictive Feature |
|---|
| Current Charges |
| Criminal History: self |
| Substance Abuse |
| Stability of Employment & Living Situation |
| Personality |
| Criminal Attitudes |
| Neighborhood Safety |
| Criminal History: family and friends |
| Quality of Social Life & Free Time |
| Education & School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

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

# *Main Survey*: Survey Samples




| Gender | AMT | SSI | Census |
|---|---|---|---|
| Male | 0.55 | 0.44 | 0.49 |
| Female | 0.43 | 0.55 | 0.51 |

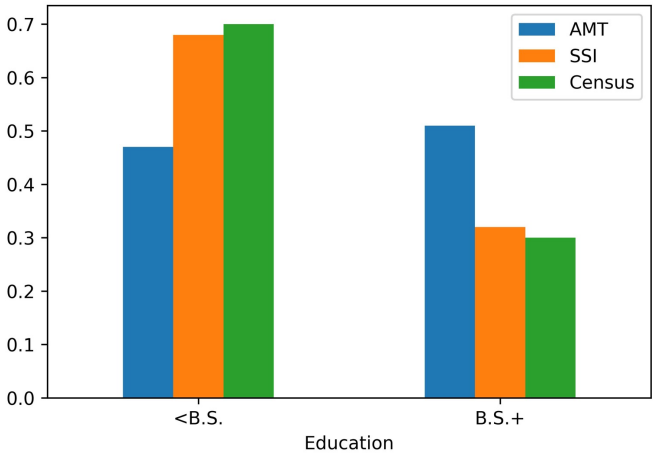



| Race | AMT | SSI | Census |
|---|---|---|---|
| African American | 0.09 | 0.12 | 0.13 |
| Asian | 0.03 | 0.04 | 0.06 |
| Caucasian | 0.76 | 0.71 | 0.61 |
| Hispanic | 0.08 | 0.11 | 0.18 |
| Other | 0.02 | 0.02 | 0.04 |

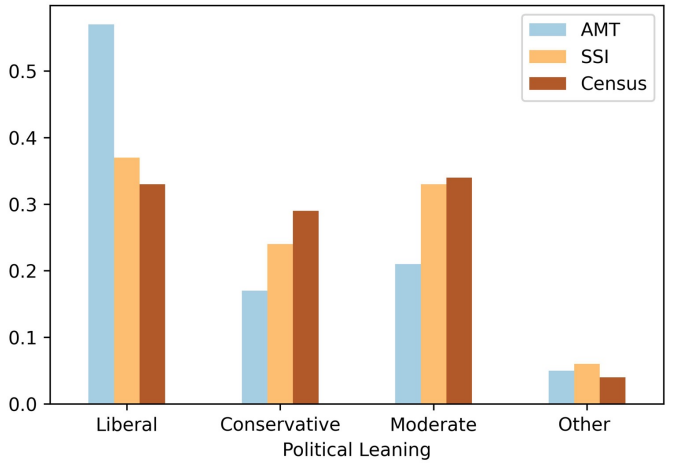
AMT (*Amazon Mechanical Turk* ) : 196
SSI (*Survey Sampling International*) : 380

Total : 576

# *Main Survey*: Survey Samples




| Education | AMT | SSI | Census |
|---|---|---|---|
| **<B.S.** | 0.47 | 0.68 | 0.7 |
| **B.S.+** | 0.51 | 0.32 | 0.3 |





| Politics | AMT | SSI | Census |
|---|---|---|---|
| **Liberal** | 0.57 | 0.37 | 0.33 |
| **Conservative** | 0.17 | 0.24 | 0.29 |
| **Moderate** | 0.21 | 0.33 | 0.34 |
| ***Other*** | 0.05 | 0.06 | 0.04 |

AMT (*Amazon Mechanical Turk* ) : 196
SSI (*Survey Sampling International*) : 380

Total : 576

# Limitations

- Self-report studies may suffer from generalizability biases, as those who take these surveys may not be representative of the general population.
  - Recruited both AMT & SSI samples to maximize generalizability and data quality.
- Self-report biases may affect the data.
  - Applied extensive pre-testing & question randomization to mitigate these self-report biases.
- People may feel differently about fairness in different contexts.
  - Future work should seek to validate whether a model that produces actual results based on self-report inputs when people are placed in more ecologically-valid situations.

# To be continued in [*Part 2*]…

- Fairness Judgements Analysis
  - Across different features
  - Across different users
- Fairness Resoning Analysis
- Latent Properly Assessments
- Modelling Fairness Judgement
- Predicting & Evaluating Fairness Judments
- Investigation on Fairness Disagreements

Thank
You