Human Perceptions of Fairness

Matheus Bustamante, Jack Phillips, Siam Abd Al-Ilah

Outline

- Algorithmic fairness
- Definition of Fairness
- Hypothesis
- Caveats
- Results
- Procedural Fairness
- Procedures
- Results

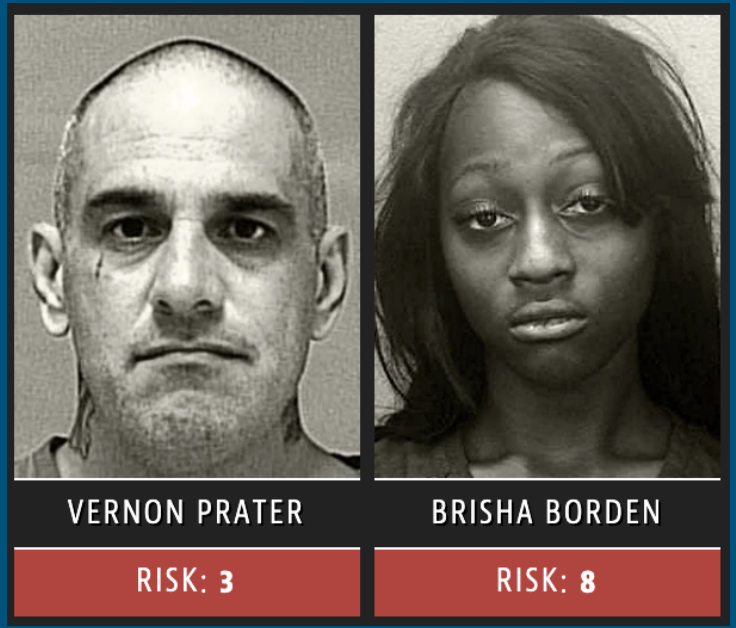
Algorithmic Fairness

- Machine Learning tools are increasingly being used to make decisions for humans.
- Given that such decisions may have a long lasting impact on people's lives, it is important to ask whether these decisions are fair.

Algorithmic Fairness

- ProPublica
  - False Positive Rate for African-Americans is much higher than Caucasians.

- Northpointe
  - Positive Predictive Value and False Omission rate is similar for African-Americans and Caucasians.



Definition of Fairness

- Demographic Parity
- Disparate Impact
- Equality of odds
- Calibration
- Counterfactual fairness
- Equality of opportunity
- Individual fairness
- Predictive parity

Link: https://developers.google.com/machine-learning/glossary/fairness

Definition of Fairness

- Demographic Parity
  - Seek to equalize the percentage of people who are predicted to be positive across different groups.

Definition of Fairness

- Error Parity
  - Requires the overall proportion of classification errors to be equal across all demographic groups

Definition of Fairness

- Equality of False Positive or False Negative rates
  - Requires the percentage of people that were wrongly predicted to be negative/positive to be the same across individuals belonging to different groups.

Definition of Fairness

- False Discovery or Omission rates
  - Equalize the percentage of false positive/negative predictions among individuals predicted to be positive/negative in each group.

Mathematical Definition of Fairness

I: Decision subject

G: Group

Yi: True Label

y^i: predicted label

n_G: Number of individuals at group G

Fairness notion	benefit for group G
DP	$b^G = \frac{1}{n_G} \sum_{i \in G} 1[\hat{y}_i = 1]$
EP	$b^G = \frac{1}{n_G} \sum_{i \in G} 1[\hat{y}_i \neq y_i]$
FDP	$b^G = \frac{\sum_{i \in G} 1[y_i=0 \& \hat{y}_i=1]}{\sum_{i \in G} 1[\hat{y}_i=1]}$
FNP	$b^G = \frac{\sum_{i \in G} 1[\hat{y}_i=0 \& y_i=1]}{\sum_{i \in G} 1[y_i=1]}$

What does the research attempt to do?

- Various mathematical formulations of fairness have been proposed.
- Current literature attempts to quantify the pros and cons of different definitions of fairness.
- In this paper, they determine what notion of fairness is most compatible with lay-people's perception of fairness in a particular context.

Discussion I

1. With coronavirus vaccines becoming ready for release, there is a big question about how we release them. What do you think is the most fair way to distribute the vaccines, and how does it relate to some of the different types of fairness we have discussed?

(5 min)

Hypotheses

H1: In the context of recidivism risk assessment, the majority of subjects' responses is compatible with equality of *false negative/positive* rates across demographic groups.

H2: In the context of medical predictions, the majority of subjects' responses is compatible with *equality of accuracy* across demographic groups.

H3: When the decision-making stakes are high (e.g., when algorithmic predictions affect people's life expectancy) participants are more sensitive to accuracy as opposed to equality.

Quick note on EC2 active learning algorithm

- Required around 600 tests to obtain likelihood for a notion of fairness

- Using EC2 (Equivalence Class Edge Cutting) algorithm allows to cut in down to 20

- On a high level, the algorithm looks at the tests already administered, and decides which test will ‘narrow down’ the equivalence classes

Demographic Parity was the most preferred notion

Table 3: Number of participants matched with each notion with high likelihood (> 80%).

	DP	EP	FDR	FOP	none
Crime risk prediction	80%	0%	2%	4%	14%
Cancer risk prediction	73%	3%	0%	0%	24%

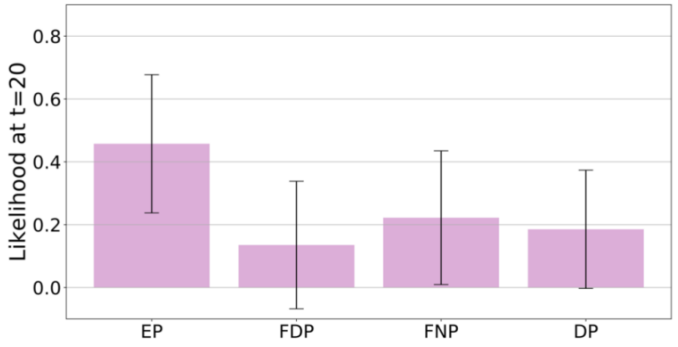


Figure 4: The outcome of our adaptive experiment if the response to each test is chosen at random. The chance of DP being selected as the most compatible hypothesis is not high.

Caveats

- Algorithms had roughly equal accuracy
- Respondents may not be very familiar with fairness definitions (is this a problem?)
- Amazon MTurk not representative of US population

Table 2: Demographics of our AMT participants compared to the 2016 U.S. census data.

Demographic Attribute	AMT	Census
Male	53%	49%
Female	47%	51%
Caucausian	68%	61%
African-American	12%	13%
Asian	10%	6%
Hispanic	6%	18%
Liberal	74%	33%
Conservative	19%	29%
High school	31%	40%
College degree	48%	48%
Graduate degree	20%	11%
18–25	14%	10%
25–40	67%	20%
40–60	16%	26%

Hypotheses

H1: In the context of recidivism risk assessment, the majority of subjects' responses is compatible with equality of *false negative/positive* rates across demographic groups.

H2: In the context of medical predictions, the majority of subjects' responses is compatible with *equality of accuracy* across demographic groups.

H3: When the decision-making stakes are high (e.g., when algorithmic predictions affect people's life expectancy) participants are more sensitive to accuracy as opposed to equality.

On High Stakes issues preference was on accuracy

Table 5: Three hypothetical algorithms offering distinct tradeoffs between accuracy and inequality.

Algorithm	accuracy	female acc.	male acc.
A_1	94%	89%	99%
A_2	91%	90%	92%
A_3	86%	86%	86%

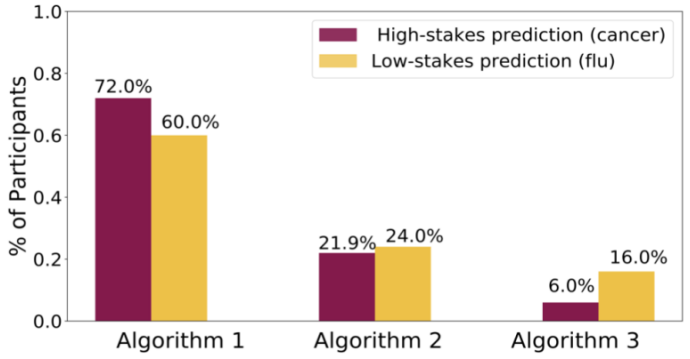


Figure 8: Medical risk prediction scenarios. Participants gave higher weight to accuracy (compared to inequality) when predictions can impact patients' life expectancy.

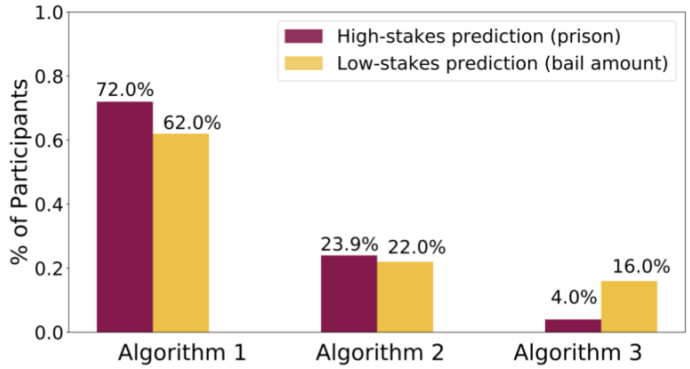


Figure 9: Crime risk prediction scenarios. Participants gave higher weight to accuracy (compared to inequality) when predictions can impact defendants' life trajectory.

Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences

Ruotong Wang
Carnegie Mellon University
ruotongw@andrew.cmu.edu

F. Maxwell Harper*
Amazon
fmh@amazon.com

Haiyi Zhu
Carnegie Mellon University
haiyiz@cs.cmu.edu

Discussion 2

- Now that we have established a few definitions on the fairness of predictions, what do you think are some other factors that can influence our perception of whether an algorithm is fair? (e.g. even if the outcomes are fair, would people judge an algorithm who uses race explicitly as an input as a fair algorithm?)

Experiment procedures

- Step 1 - Modify hypothetical algorithm's creation and deployment in 4 categories: transparency (high vs low), design (CS team, outsourced, CS+HR), model (ML vs rules), and decision (algorithm only vs mixed)
- Step 2 - Ask MTurker what the outcome will be on their profile
- Step 3 - Give MTurker randomly generated outcome
- Step 4 - Ask survey questions including demographics and perception of fairness of the algorithm
- Step 5 - Debrief

Algorithm Outcomes

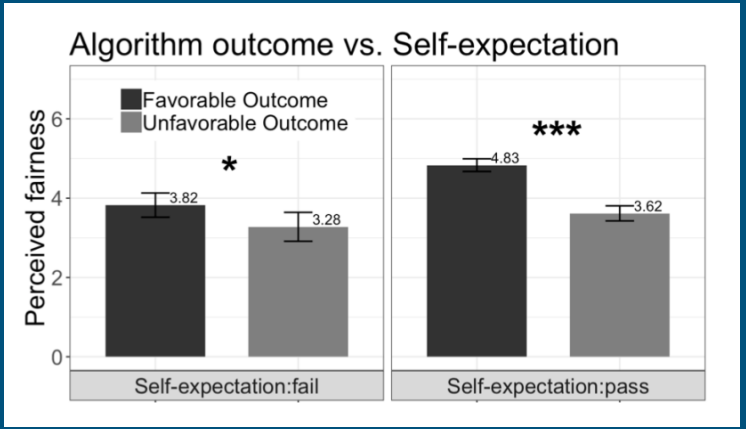
- Hypothesis 1a - *People who receive a favorable outcome think the algorithm is more fair than people who receive an unfavorable outcome.*

- Hypothesis 1b - *People perceive algorithms that are not biased against particular subgroups as more fair.*

- Hypothesis 1c - *In the context of algorithmic decision-making, the effect of a favorable outcome on perceived fairness is larger than the effect of being not biased against particular groups.*

Algorithm Outcomes - Results

	Perceived Fairness	
	Model 1	Model 2
	Coef. (S.E.)	Coef. (S.E.)
Unfavorable Outcome vs. Favorable Outcome	−1.041*** (0.113)	−0.887*** (0.258)
Biased Treatment vs. Unbiased Treatment	−0.395*** (0.113)	−1.068*** (0.257)
CS Team vs. Outsourced	0.131 (0.138)	−0.614* (0.240)
CS and HR vs. Outsourced	0.201 (0.137)	−0.352 (0.224)
Machine Learning vs. Rules	0.138 (0.112)	0.086 (0.189)
High Transparency vs. Low Transparency	−0.154 (0.114)	0.056 (0.191)
Mixed Decision vs. Algorithm-only	0.090 (0.113)	0.220 (0.189)
Unfavorable Outcome × CS team		0.426 (0.272)
Unfavorable Outcome × CS and HR		0.015 (0.273)
Unfavorable Outcome × Machine Learning		−0.238 (0.220)
Unfavorable Outcome × High Transparency		0.106 (0.224)
Unfavorable Outcome × Mixed Decision		−0.420 (0.221)
Biased Treatment × CS team		0.971*** (0.272)
Biased Treatment × CS and HR		1.070*** (0.271)
Biased Treatment × Machine Learning		0.343 (0.220)
Biased Treatment × High Transparency		−0.486* (0.224)
Biased Treatment × Mixed Decision		0.174 (0.221)
Self-expected Pass vs. Self-expected Fail	0.659*** (0.127)	0.616*** (0.125)
Constant	4.131*** (0.178)	4.438*** (0.235)
R^2	0.190	0.235
Adjusted R^2	0.178	0.210
Note:		*p<0.05; **p<0.01; ***p<0.001



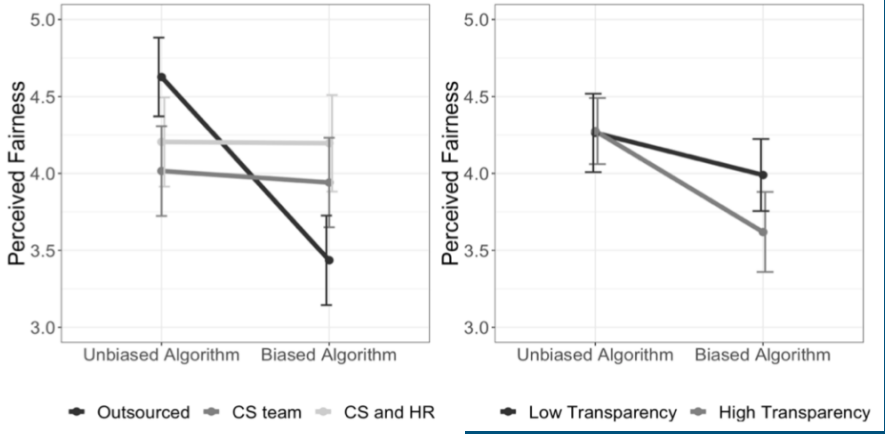
Algorithm Development Procedures

- Hypothesis 2a - *An algorithmic decision-making process that is more transparent is perceived as more fair than a process that is less transparent.*

- Hypothesis 2b - *An algorithmic decision-making process that has more human involvement is perceived as more fair than a process that has less human involvement*

Algorithm Development Procedures - Results

	Perceived Fairness	
	Model 1	Model 2
	Coef. (S.E.)	Coef. (S.E.)
Unfavorable Outcome vs. Favorable Outcome	−1.041*** (0.113)	−0.887*** (0.258)
Biased Treatment vs. Unbiased Treatment	−0.395*** (0.113)	−1.068*** (0.257)
CS Team vs. Outsourced	0.131 (0.138)	−0.614* (0.240)
CS and HR vs. Outsourced	0.201 (0.137)	−0.352 (0.224)
Machine Learning vs. Rules	0.138 (0.112)	0.086 (0.189)
High Transparency vs. Low Transparency	−0.154 (0.114)	0.056 (0.191)
Mixed Decision vs. Algorithm-only	0.090 (0.113)	0.220 (0.189)
Unfavorable Outcome × CS team		0.426 (0.272)
Unfavorable Outcome × CS and HR		
Unfavorable Outcome × Machine Learning		
Unfavorable Outcome × High Transparency		
Unfavorable Outcome × Mixed Decision		
Biased Treatment × CS team		
Biased Treatment × CS and HR		
Biased Treatment × Machine Learning		
Biased Treatment × High Transparency		
Biased Treatment × Mixed Decision		
Self-expected Pass vs. Self-expected Fail		
Constant		
R^2		
Adjusted R^2		
Note:		



Individual Differences

- Hypothesis 3a - *People with a higher level of education will perceive algorithmic decision-making to be more fair than people with a lower level of education*

- Hypothesis 3b - *People with high computer literacy will perceive algorithmic decision-making to be more fair than people with low computer literacy.*

- Hypothesis 3c - *People in demographic groups that typically benefit from algorithmic biases (young, white, men) will perceive algorithmic decision-making to be more fair than people in other demographic groups (old, non-white, women)*

Individual Differences - Results