Measuring Non-Expert Comprehension of Machine Learning Fairness Metrics

Debjani Saha [1] **Candice Schumann** [1] **Duncan C. McElfresh** [1] **John P. Dickerson** [1] **Michelle L. Mazurek** [1] **Michael Carl Tschantz** [2]

Abstract

Bias in machine learning has manifested injustice in several areas, such as medicine, hiring, and criminal justice. In response, computer scientists have developed myriad definitions of *fairness* to correct this bias in fielded algorithms. While some definitions are based on established legal and ethical norms, others are largely mathematical. It is unclear whether the general public agrees with these fairness definitions, and perhaps more importantly, whether they *understand* these definitions. We take initial steps toward bridging this gap between ML researchers and the public, by addressing the question: *does a lay audience understand a basic definition of ML fairness?* We develop a metric to measure comprehension of three such definitions–demographic parity, equal opportunity, and equalized odds. We evaluate this metric using an online survey, and investigate the relationship between comprehension and sentiment, demographics, and the definition itself.

1. Introduction

Research into algorithmic fairness has grown in both importance and volume over the past few years, driven in part by the emergence of a grassroots Fairness, Accountability, Transparency, and Ethics (FATE) in Machine Learning (ML) community. Different metrics and approaches to algorithmic fairness have been proposed, many of which are based on prior legal and philosophical concepts, such as disparate impact and disparate treatment (Feldman et al., 2015; Chouldechova, 2017; Binns, 2017). However, definitions of ML fairness do not always fit well within pre-existing legal and moral frameworks. The rapid expansion of this field makes it difficult for professionals to keep up, let alone the general public. Furthermore, misinformation about notions of fairness can have significant legal implications.[1]

Computer scientists have largely focused on developing mathematical notions of fairness and incorporating them into ML systems. A much smaller collection of studies have measured public perception of bias and (un)fairness in algorithmic decision-making. However, as both the academic community and society in general continue to discuss issues of ML fairness, it remains unclear whether non-experts–who will be *impacted* by ML-guided decisions–understand various mathematical definitions of fairness sufficiently to provide opinions and critiques. We emphasize that these technologies are likely to have greater impact on marginalized populations, and those with lower levels of education, as in the case of hiring and criminal justice (Barocas & Selbst, 2016; Frey & Osborne, 2017). For this reason, we focus on a non-expert audience and a context (hiring) that most people would find relatively familiar.

Our Contributions. We take a step toward addressing this issue by studying peoples' comprehension and perceptions of three definitions of ML fairness: *demographic parity*, *equal opportunity,* and *equalized odds* (Hardt et al., 2016). Specifically, we address the following research questions:

- **RQ1** When provided with an explanation intended for a non-technical audience, do non-experts comprehend each definition and its implications?
- **RQ2** What factors play a role in comprehension?
- **RQ3** How are comprehension and sentiment related?
- **RQ4** How do the different definitions compare in terms of comprehension?

We developed two online surveys to address these research questions. We presented participants with a simplified decision-making scenario and an accompanying *fairness rule* expressed in the scenario's context. We asked questions related to the participants' comprehension of and sentiment toward this rule. Tallying the number of correct responses to the comprehension questions gives us a *comprehension score* for each participant. In Study-1, we found that this comprehension score is a consistent and reliable indicator of understanding demographic parity.

[1]University of Maryland, College Park, MD [2]ICSI, Berkeley, CA. Correspondence to: Michelle L. Mazurek <mmazurek@cs.umd.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1]https://www.cato.org/blog/misleading-veritas-accusation-google-bias-could-result-bad-law

Then, in Study-2, we used a similar approach to compare comprehension among all three definitions of interest. We find that (1) education is a significant predictor of rule understanding, (2) the counterintuitive definition of Equal Opportunity with False Negative Rate was significantly harder to understand than other definitions, and (3) participants with low comprehension scores tended to express less negative sentiment toward the fairness rule. This underlines the importance of considering stakeholders before deploying a "fair" ML system, because some stakeholders may not understand or agree with an ML-specific notion of fairness. Our goal is to help to designers and adopters of fairness approaches understand whether they are communicating with stakeholders effectively.

2. Related Work

In response to many instances of bias in fielded artificial intelligence (AI) and machine learning (ML) systems, ML fairness has received significant attention from the computer-science community. Notable examples include gender bias in job-related ads (Datta et al., 2015), racial bias in evaluating names on resumes (Caliskan et al., 2017), and racial bias in predicting criminal recidivism (Angwin et al., 2016). To correct biased behavior, researchers have proposed several mathematical and algorithmic notions of fairness.

Most algorithmic fairness definitions found in literature are motivated by the philosophical notion of individual fairness (e.g., see (Rawls, 1971)), and legal definitions of disparate impact/treatment (e.g., see (Barocas & Selbst, 2016)). Several ML-specific definitions of fairness have been proposed which claim to uphold these philosophical and legal concepts. These definitions of "ML fairness" fall loosely into two categories (for a review, see (Chouldechova & Roth, 2018)). *Statistical Parity* posits that in a *fair* outcome, individuals from different protected groups have the same chance of receiving a positive (or negative) outcome. Similarly, *Predictive Parity* (Hardt et al., 2016) asserts that the predictive accuracy should be similar across different protected groups–often measured by the false positive rate (FPR) or false negative rate (FNR) in binary classification settings. Myriad other definitions have been proposed, based on concepts such as calibration (Pleiss et al., 2017) and causality (Kusner et al., 2017). Of course, all of these definitions make limiting assumptions; no concept of fairness is perfect (Hardt et al., 2016). The question remains, *which* of these fairness definitions are appropriate, and in *what context?* There are two important components to answering this question: *communicating* these fairness definitions to a general audience, and *measuring their perception* of these definitions in context.

Communicating ML-related concepts is an active and growing research area. In particular, *interpretable ML* focuses on communicating the decision-making process and results of ML-based decisions to a general audience (Lipton, 2018). Many tools have been developed to make ML models more interpretable, and many demonstrably improve understanding of ML-based decisions (Ribeiro et al., 2016; Huysmans et al., 2011). These models often rely on concepts from probability and statistics—teaching these concepts has long been an active area of research. Batanero et al. (2016) provide an overview of teaching probability and how students learn probability; our surveys use their method of communicating probability, which relies on proportions. We draw on several other concepts from this literature for our study design; for example avoiding numerical and statistical representations (Gigerenzer & Edwards, 2003; Gigerenzer et al., 2007), which can be confusing to a general audience. Instead we provide relatable examples, accompanied by examples and graphics (Hogarth & Soyer, 2015).

Effectively communicating ML concepts is necessary to achieve our second goal of understanding peoples' perceptions of these concepts. One particularly active research area focuses on how people perceive bias in algorithmic systems. For example, Woodruff et al. (2018) investigated perceptions of algorithmic bias among marginalized populations, using a focus group-style workshop; Grgic-Hlaca et al. (2018) study the underlying factors causing perceptions of bias, highlighting the importance of selecting appropriate features in algorithmic decision-making; Plane et al. (2017) look at perceptions of discrimination of online advertising; Harrison et al. (2020) studies perceptions of fairness in stylized machine learning models; Srivastava et al. (2019) note that perceived appropriateness of an ML notion of fairness may depend on the domain in which the decision-making system is deployed, but suggest that simpler notions may best capture lay perceptions of fairness.

A related body of work studied how people perceive algorithmic decision-makers. Lee (2018) studies perceptions of fairness, trust, and emotional response of algorithmic decision-makers — as compared to human decision-makers. Similar work studies perception of fairness in the context of splitting goods or tasks, and in loan decisions (Lee & Baykal, 2017; Lee et al., 2019; Saxena et al., 2020). Binns et al. (2018) studies how different explanation styles impact perceptions of algorithmic decision-makers.

This substantial body of prior research provided inspiration and guidance for our work. Prior work has studied both the effective communication of, and perceptions of, ML-related concepts. We hypothesize that these concepts are in fact related; to that end, we design experiments to simultaneously study peoples' *comprehension* of and *perceptions* of common ML fairness definitions.

3. Methods

To study perceptions of ML fairness, we conducted two online surveys where participants were presented with a hypothetical decision-making scenario. Participants were then presented with a "rule" for enforcing fairness. We then asked each participant several questions on their comprehension and perceptions of this fairness rule. We first conducted Study-1 to validate our methodology; we then conducted the larger and broader Study-2 to address our main research questions. Both studies were approved by the University of Maryland Institutional Review Board (IRB).

3.1. Study-1

In Study-1 we tested three different decision-making scenarios based on real-world decision problems: hiring, giving employee awards, and judging a student art project. However, we observed no difference in participant responses between these scenarios; for this reason, we focus exclusively on hiring in Study-2 (see 3.2). Please see Appendix D for a description of the Study-1 scenarios, and Appendix B.5 for relevant survey results. In Study-1, we chose (what we believe is) the simplest definition of ML fairness, namely, demographic parity. In short, this rule requires that the fraction of one group who receives a *positive* outcome (e.g., an award or job offer) is equal for both groups.

3.1.1. Survey Design

Here we provide a high-level discussion of the survey design; the full text of each survey can be found in Appendix D. The participant first receives a consent form (see Appendix E). If consent is obtained, the participant sees a short paragraph explaining the decision-making scenario. To make demographic parity accessible to a non-technical audience, and to avoid bias related to algorithmic decision-making, we frame this notion of fairness as a *rule* that the decision-maker must follow to be fair. In the hiring scenario, we framed this decision rule as follows: *The fraction of applicants who receive job offers that are female should equal the fraction of applicants that are female. Similarly, the fraction of applicants who receive job offers that are male should equal the fraction of applicants that are male.*

We then ask two questions concerning participant evaluation of the scenario, nine comprehension questions about the fairness rule, two self-report questions on participant understanding and use of the rule, and four free response questions on comprehension and sentiment. For example, one comprehension question is: *Is the following statement TRUE OR FALSE: This hiring rule always allows the hiring manager to send offers exclusively to the most qualified applicants*. Finally, we collect demographic information (age, gender, race/ethnicity, education level, and expertise in a number of relevant fields).

We conducted in-person cognitive interviews (Harrell & Bradley, 2009) to pilot our survey, leading to several improvements in the question design. Most notably, because some cognitive interview participants appeared to use their own personal notions of fairness rather than our provided rule, we added questions to assess this compliance issue.

3.1.2. Recruitment and Participants

We recruited participants using the online service Cint (Cint), which allowed us to loosely approximate the 2017 U.S. Census distributions (Bureau, 2017) for ethnicity and education level, allowing for broad representation. We required that participants be 18 years of age or older, and fluent in English. Participants were compensated using Cint's rewards system; according to a Cint representative: "[Participants] can choose to receive their rewards in cash sent to their bank accounts (e.g. via PayPal), online shopping opportunities with one of multiple online merchants, or donations to a charity."

Data was collected during August 2019. In total 147 participants were included in the Study-1 analysis, including 75 men (51.0%), 71 women (48.3%), and 1 (0.7%) preferring not to answer. The average age was 46 years (SD = 16). Ethnicity and educational attainment are summarized in Table 1. On average, participants completed the survey in 14 minutes.

Table 1 summarizes the ethnicity and education level of participants in both Study-1 and Study-2.

Table 1. Participant demographics across ethnicity and education level, compared to the 2017 U.S. Census. AI = American Indian, AN = Alaska Native, NH = Native Hawaiian, PI = Pacific Islander, AA = African American. Note that in Study-2, two participants did not report their education level.

	Percent of Sample		
	Census	**Study-1**	**Study-2**
Ethnicity			
AI or AN	0.7	0.7	0.9
Asian or NH or PI	5.7	1.4	2.3
Black or AA	12.3	10.2	15.8
Hispanic or Latinx	18.1	12.2	7.7
Other	2.6	2.7	1.4
White	60.6	72.8	71.9
Education Level			
Less than HS	12.1	6.1	6.9
HS or equivalent	27.7	29.9	24.9
Some post-secondary	30.8	30.6	24.9
Bachelor's and above	29.4	33.3	42.7

3.2. Study-2

Study-2 follows a very similar structure to Study-1 with a few changes. First, we decided to use only the hiring (HR) decision scenario (See Appendix B.5 for more in-depth discussion). Second, we expanded to three definitions of fairness: *demographic parity* (DP), *equal opportunity* (EP), and *equalized odds* (EO) (Hardt et al., 2016). Within EP, we tested both False Negative Rate (FNR) and False Positive Rate (FPR), resulting in a total of four conditions.

3.2.1. Survey Design

Here we provide a high-level discussion of the differences between Study-2 and Study-1; the full text of each survey can be found in Appendix D. We used a between-subjects design with random assignment among the four conditions (DP, FNR, FPR, EO). Again, we frame each notion of fairness as a *hiring rule* that the decision-maker must follow to be fair. For example, in FPR we define the award rule as follows: *The fraction of unqualified male candidates who receive job offers should equal the fraction of unqualified female candidates who receive job offers.*

For this version, we added graphical examples to further clarify our explanations (see Fig. 1 for an example). We used the all the same questions as in Study-1 but added two additional Likert-scale questions assessing participant sentiment: one asked whether they liked the rule, and the other asked whether they agreed with the rule. One free response question (asking how participants personally would go about the hiring process to ensure it was fair), which did not consistently provide useful responses in Study-1, was removed from the Study-2 survey in an effort to keep the expected completion time similar.

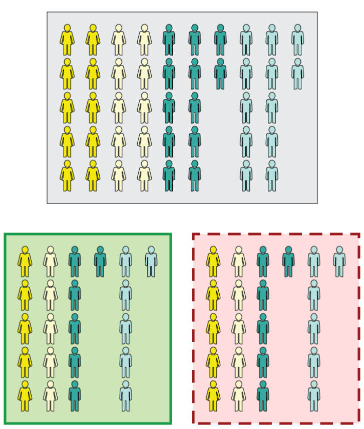


Figure 1. A graphical example to describe a fair hiring outcome for EO. Yellow people represent females while green people represent males. The darker colors represent qualified individuals while the lighter colors represent unqualified individuals. The gray box represents the original pool of applicants. The green box represent individuals that received job offers while the red box with a dashed border represents individuals that did *not* receive job offers.

3.2.2. Recruitment and Participants

We again used the Cint service to recruit participants. Compensation for participation was handled in the same manner as described in §3.1.2. Because our initial sample (intended to target education, ethnicity, gender and age distributions approximating the U.S. census) skewed more highly educated than we had hoped, we added a second round of recruitment one week later primarily targeting participants without bachelor's degrees. Hereafter, we report on both samples together.

Data was collected during January and February 2020. In total 349 participants were included in the Study-2 analysis, including 142 men (40.7%), 203 women (58.2%), 1 other (0.3%), and 3 (0.9%) preferring not to answer. The average age was 45 years (SD = 15). Ethnicity and educational attainment are summarized in Table 1. On average, participants completed the survey in 16 minutes.

3.3. Data Analysis

Free response questions were qualitatively coded for statistical testing. In Study-1, one question was coded by a single researcher for simple correctness (see Appendix B.1), and the other was independently coded by three researchers (resolved to 100%) to capture sentiment information (see Appendix B.3). In Study-2, both questions were independently coded by 2-3 researchers (resolved to 100%). Participants who provided nonsensical answers, answers not in English, or other non-responsive answers to free response questions were excluded from all analysis.

The following methods were used for all statistical analyses unless otherwise specified. Correlations with nonparamentric ordinal data were assessed using Spearman's rho. Omnibus comparisons on nonparametric ordinal data were performed with a Kruskal–Wallis (K-W) test, and relevant post-hoc comparisons with Mann–Whitney U (M-WU) tests. Post-hoc p-values were adjusted for multiple comparisons using Bonferroni correction. χ^2 tests were used for comparisons of nominal data. Boxplots show median and first and third quartiles; whiskers extend to $1.5 * \text{IQR}$ (interquartile range), with outliers indicated by points. The full analysis script for both studies can be found on github. [2]

3.4. Limitations

As with all surveys, our study has certain limitations. We recruited a demographically broad population, but web panels are generally more tech-savvy than the broader population (Redmiles et al., 2019). We consider this acceptable for a first effort. Some participants may be satisficing rather

[2] https://github.com/saharaja/ICML2020-fairness

than answering carefully. We mitigate this by disqualifying participants with off-topic or non-responsive free-text responses. Further, this limitation can be expected to be consistent across conditions, enabling reasonable comparison. Finally, better or clearer explanations of the fairness definitions we explored are certainly possible; we believe our explanations were sufficient to allow us to investigate our research questions, especially because they were designed to be consistent across conditions.

4. Results

In this section we first discuss the preliminary findings from Study-1 (see §4.1). These findings were used as hypotheses for further exploration and testing in Study-2; we discuss those results second (see §4.2).

4.1. Study-1

We analyze survey responses for Study-1 and make several observations. We first validate our comprehension score as a measure of participant understanding; we then generate hypotheses for further exploration in Study-2.

4.1.1. Our Survey Effectively Captures Rule Comprehension

We find that we can measure comprehension of the fairness rule. The comprehension score was calculated as the total correct responses out of a possible 9. All questions were weighted equally. The relevant questions included 2 multiple choice, 4 true/false, and 3 yes/no questions. The average score was 6.2 (SD=2.3).

We validate our comprehension score using two methods: internal validity testing, and correlation against two self-report and one free response question included in our survey (see Appendix B.1 for further details).

Internal Validity Cronbach's α and item-total correlation were used to assess internal validity of the comprehension score. Both measures met established thresholds (Nunnally, 1978; Everitt & Skrondal, 2010): Cronbach's $\alpha = 0.71$, and item-total correlation for 8 of the 9 items (all but Q5) > 0.3.

Question Correlation We find that self-reported rule understanding and use are reflected in comprehension score. First, we compared comprehension score to self-reported rule understanding (Q13): "I am confident I know how to apply the award rule described above," rated on a five-point Likert scale from strongly agree (1) to strongly disagree (5). The median response was "agree" (Q1 $= 1$, Q3 $= 3$). Higher comprehension scores tended to be associated with greater confidence in understanding (Spearman's $\rho = 0.39$, $p < 0.001$), supporting the notion that comprehension score is a valid measure of rule comprehension.

Next, we compared comprehension score to a self-report question about the participant's use of the rule (Q14), with the following options: (a) "I applied the provided award rule only," (b) "I used my own ideas of what the correct award decision should be rather than the provided award rule," or (c) "I used a combination of the provided award rule and my own ideas of what the correct award decision should be." We find that participants who claimed to use only the rule scored significantly higher (mean 7.09) than those who used their own notions (4.90) or a combination (4.68) (post-hoc M-WU, $p < 0.001$ for both tests; corrected $\alpha = 0.05/3 = 0.017$). This further corroborates our comprehension score.

Finally, we asked participants to explain the rule in their own words (Q12). Each response was then qualitatively coded as one of five categories – **Correct**: describes rule correctly; **Partially correct**: description has some errors or is somewhat vague; **Neither**: vague description of purpose of the rule rather than how it works, or pure opinion; **Incorrect**: incorrect or irrelevant; and **None**: no answer, or expresses confusion. Participants whose responses were either correct (mean comprehension score = 7.71) or partially correct (7.03) performed significantly better on our survey than those responding with neither (5.13) or incorrect (4.24) (post-hoc M-WU, $p < 0.001$ for these four comparisons, corrected $\alpha = 0.05/10 = 0.005$). These findings further validate our comprehension score. Additional details of these results and the associated statistical tests can be found in Appendix B.1.

4.1.2. Hypotheses Generated

We analyzed the data from Study-1 in an exploratory fashion intended to generate hypotheses that could be tested in Study-2. We highlight here three key hypotheses that emerged from the data.

Education Influences Comprehension We used poisson regression models to explore whether various demographic factors were associated with differences in comprehension. We found that a model including education as a regressor had greater explanatory power than a model without (see Appendix B.2 for further details).

Disagreement with the Rule is Associated with Higher Comprehension Scores We asked participants for their opinion on the presented rule in a free response question (Q15). These responses were qualitatively coded to capture participant sentiment toward the rule in one of five categories – **Agree**: generally positive sentiment towards rule; **Depends**: describes both pros and cons of the given rule; **Disagree**: generally negative sentiment towards rule; **Not understood**: expresses confusion about rule; **None**: no answer, or lacks opinion on appropriateness of the rule. Participants who expressed disagreement with the rule performed better (mean comprehension score = 7.02) than those who expressed agreement (5.50), did not understand the rule (4.44), or provided no response (5.09) to the question (post-hoc M-WU, $p < 0.005$ for these three comparisons; corrected $\alpha = 0.05/10 = 0.005$). Appendix B.3 provides further details.

Non-Compliance is Associated with Lack of Understanding We were interested in understanding why some participants failed to adhere to the rule, as measured by their self-report of rule usage in Q14. We labeled those who responded with either having used their own personal notions of fairness ($n = 29$) or some combination of their personal notions and the rule ($n = 28$) as "non-compliant" (NC), with the remaining $n = 89$ labeled as "compliant" (C). One participant who did not provide a response was excluded from this analysis, conducted using χ^2 tests.

Non-compliant participants were less likely to self-report high understanding of the rule in Q13 (see Fig. 13). Moreover, non-compliance also appears to be associated with a reduced ability to correctly explain the rule in Q12 (see Fig. 14). This fits with the overall strong relationship we observed among comprehension scores, self-reported understanding, ability to explain the rule, and compliance.

Further, negative participant sentiment towards the rule (Q15) also appears to be associated with greater compliance (see Fig. 15). Thus, non-compliant participants appear to behave this way because they do not *understand* the rule, rather than because they do not *like* it. Refer to Appendix B.4 for further details.

4.2. Study-2

We first confirm the validity of our comprehension score, then compare comprehension across definitions and examine the hypotheses generated in Study-1.

4.2.1. Score Validation

We validated our metric using the same approach used in Study-1, i.e., assessing both internal validity and correlation with self-report and free-response questions. We report the results of this assessment here.

Internal Validity We again used Cronbach's α and item-total correlation to assess internal validity of the comprehension score. An initial assessment using all 349 responses yielded Cronbach's $\alpha = 0.38$, and item-total correlation > 0.3 for only four of the nine comprehension questions. Since both measures performed below established thresholds (Nunnally, 1978; Everitt & Skrondal, 2010), we investigated further and repeated these measurements individually for each fairness-definition condition (DP, FNR, FPR, EO). This procedure showed stark differences in Cronbach's α based on definition: DP = 0.64, FNR = 0.39, FPR = 0.49, EO = 0.62. Item-total correlations followed a similar pattern: best in DP, worst in FNR. Based on these differences, we iteratively removed problematic questions from the score on a per-definition basis until all remaining questions achieved an item-total correlation of > 0.3 (Everitt & Skrondal, 2010). By removing poorly performing questions, we increase our confidence that the measured comprehension scores are meaningful for further analysis. Table 2 specifies which questions were retained for analysis in each definition.

Table 2. Questions that were used for downstream analysis after iterative removal of questions with poor item-total correlation.

	Questions								
	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
DP	X	X			X	X	X	X	X
FNR	X	X	X			X			
FPR	X	X	X	X		X		X	X
EO	X	X	X		X	X	X	X	X

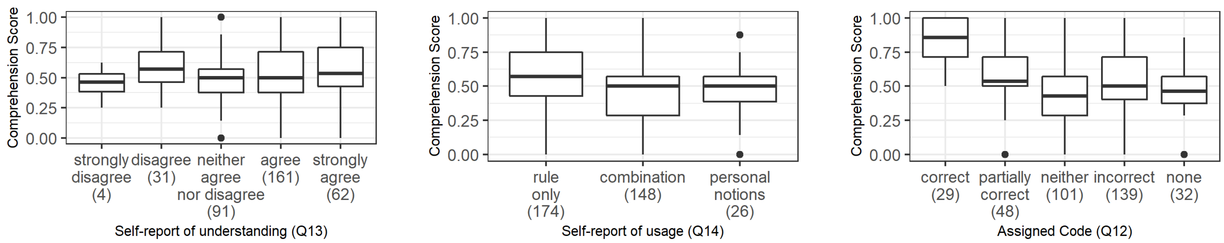
Because questions were dropped on a per-definition basis, the maximum of the resulting scores varied from 4-8 depending on the definition, rather than being a uniform 9. We normalized this treating comprehension score as a percentage of the maximum for each condition rather than a raw score. We report this *adjusted score* in the remainder of §4.2. The average score was 0.53 (SD=0.22).

Question Correlation As in Study-1, we compare comprehension scores with responses to self-report and free response questions included in our survey.

First, we compared comprehension score to self-reported rule understanding (Q13), as described in §4.1.1. The median response was "agree" (Q1 $= 2$, Q3 $= 3$). We assess the correlation between these responses and comprehension score using Spearman's rho (appropriate for ordinal data). Unlike in Study-1, there was no relationship between self-reported understanding and comprehension score (Fig. 2a).

Next, we compared comprehension score to a self-report question about the participant's use of the rule (Q14), as described in §4.1.1. A K-W test revealed a relationship between self-reported rule usage and comprehension score ($p < 0.001$). We find that participants who claimed to use only the rule tended to score higher (mean comprehension score = 0.58) than those who used a combination of the rule and their own notions of fairness (0.47, $p < 0.01$). No other differences were found (post-hoc M-WU; corrected $\alpha = 0.05/3 = 0.017$). This suggests that participants are answering at least somewhat honestly: when they try to apply the rule, comprehension scores improve (see Fig. 2b).

Finally, we asked participants to explain the rule in their own words (Q12). Each response was then qualitatively coded as



(a) Grouped by response to Q13 (b) Grouped by response to Q14. (c) Grouped by coded response to Q12.

Figure 2. Comprehension scores grouped by questions. In (a), self-reported understanding of the rule was not related to comprehension score. X-axis is reversed for figure and correlation test. In (b), rule compliance (leftmost on the x-axis) was associated with higher comprehension scores. One participant who did not provide a response was excluded from this figure and the relevant analysis. Finally, in (c), participants who provided either correct or partially correct responses tended to perform better.

one of five categories, as described in §4.1.1. These results can be seen in Fig. 2c. A K-W test revealed a relationship between comprehension score and coded responses to Q12 ($p < 0.001$). Correct (mean comprehension score = 0.83) responses were associated with higher comprehension scores than partially correct (0.58), neither (0.44), incorrect (0.52), and none (0.48) responses ($p < 0.001$ for all); partially correct responses were also associated with higher comprehension scores than neither responses ($p < 0.001$); and incorrect responses were associated with higher comprehension scores that neither responses ($p < 0.005$). No other differences were found (post-hoc M-WU; corrected $\alpha = 0.05/10 = 0.005$). These findings support our claim that our comprehension score is a valid measure of fairness-rule comprehension.

4.2.2. Education and Definition are Related to Comprehension Score

One hypothesis generated by Study-1 was that comprehension score is positively correlated with education level. We investigated this hypothesis further in Study-2 using linear regression models followed by model selection. We believe this exploratory approach to be appropriate despite the previously formulated hypothesis, given the introduction of a new variable in Study-2, i.e. fairness definition.

Eleven models were tested, regressing different combinations of demographics (ethnicity, gender, education, and age) and condition (fairness definition). Models were compared using Akaike information criterion (AIC), a standard method of evaluating model quality and performing model selection (Akaike, 1974). Comparison by AIC revealed that the model using just education (edu) and fairness definition (def) as regressors was the model of best fit. In this model, having a Bachelor's degree or above resulted in a score increase of 0.14, and the FNR condition caused a score decrease of -0.11 ($p < 0.004$ for both; corrected $\alpha = 0.05/11 = 0.0045$). A regression table of the best fit model can be found in Table 3.

Table 3. Regression table for the best fit model, with two covariates: education (baseline: no HS) and definition (baseline: DP). Est. = estimate, CI = confidence interval.

Covariate	**Est.**	**95% CI**	p
Education			
HS	0.00	[-0.10, 0.10]	0.989
Post-secondary, no BS	0.09	[-0.01, 0.18]	0.078
Bachelor's and above	0.14	[0.04, 0.23]	< 0.004
Definition			
EO	-0.08	[-0.14, 0.01]	0.020
FPR	-0.05	[-0.11, 0.01]	0.124
FNR	-0.11	[-0.18, -0.05]	< 0.001

AIC results of each of the eleven models, along with the relevant regressors, can be seen in Table 4 in Appendix C.1. Comprehension score as a function of education and fairness definition can be seen in Figs. 3 and 4.

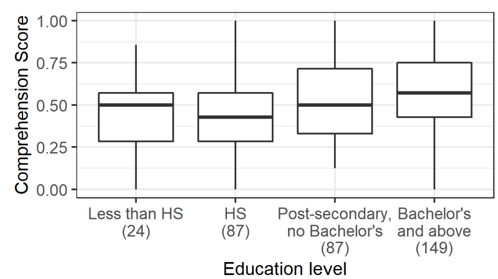


Figure 3. Comprehension score grouped by education level. Higher education was associated with higher comprehension scores. Note that two participants who did not report their education level were removed from this figure and the relevant analysis.

4.2.3. Greater Negative Sentiment Toward the Rule is Associated with Higher Comprehension Scores

In Study-1, we found a relationship between participant sentiment towards the rule and comprehension score. To better interrogate this phenomenon, in Study-2 we added two more questions to the survey to directly address the

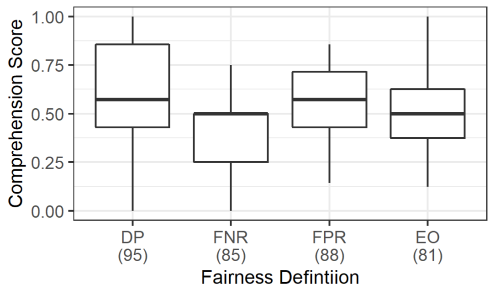


Figure 4. Comprehension score grouped by fairness definition. The FNR condition was associated with lower comprehension sore.

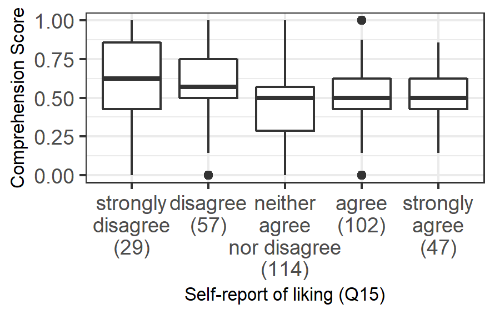


Figure 5. Comprehension score grouped by response to Q15. Dislike of the rule was associated with higher comprehension scores. X-axis is reversed for figure and correlation test.

issue of sentiment, rather than relying on a free-response question. One (Q15) asks, "To what extent do you agree with the following statement: I like the hiring rule?", and is evaluated on a five-point Likert scale from "strongly agree" (1) to "strongly disagree" (5). The other (Q16) asks, "To what extent do you agree with the following statement: I agree with the hiring rule?", and is also evaluated on a five-point Likert scale from "strongly agree" (1) to "strongly disagree" (5).

Using Spearman's rho, we assessed the correlation between responses to these two questions and comprehension score. A minor correlation was found between liking the rule and comprehension score, i.e. those who disliked the rule were more likely to have higher comprehension scores ($\rho = -0.11, p < 0.05$; see Fig. 5). A slight correlation was also found between agreeing with the rule and comprehension score, i.e. disagreement was associated with higher comprehension scores ($\rho = -0.11, p < 0.05$; see Fig. 6).

4.2.4. Non-Compliance is Associated with Lack of Understanding

A final hypothesis generated in Study-1 involves non-compliance: i.e., why do participants who report *not* using the rule to answer the comprehension questions behave this way? In Study-1, we found that this was due to the fact that non-compliant participants were less able to *understand* the rule, rather than because they did not *like* it. We also

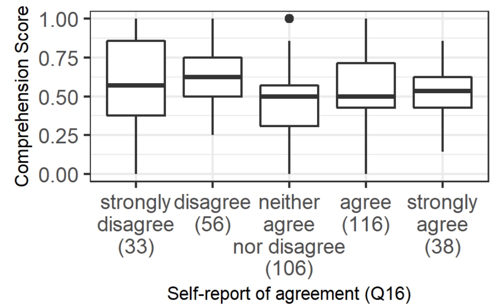


Figure 6. Comprehension score grouped by response to Q16. Disagreement with the rule was associated with higher comprehension score. X-axis is reversed for figure and correlation test.

observed this in our results from Study-2: compliant participants exhibited higher self-reported understanding of the rule ($p < 0.001$, Fig. 17), were more likely to correctly explain the rule ($p < 0.001$, Fig. 18), and were more likely to dislike the rule ($p < 0.05$, Fig. 19). We observed no relationship between compliance and agreement with the rule (Fig. 20). Refer to Appendix C.2 for more details.

5. Discussion

Bias in machine learning is a growing threat to justice; to date, ML bias has been documented in both commercial and government applications, in sectors such as medicine, criminal justice, and employment. In response, ML researchers have proposed various notions of *fairness* to correct these biases. Most ML fairness definitions are purely mathematical, and require some knowledge of machine learning. While they are intended to benefit the general public, it is unclear whether the general public agrees with — or even understands — these notions of ML fairness.

We take an initial step to bridge this gap by asking *do people understand the notions of fairness put forth by ML researchers?* To answer this question we develop a short questionnaire to assess understanding of three particular notions of ML fairness (demographic parity, equal opportunity, and equalized odds). We find that our comprehension score (with some adjustments for each definition) appears to be a consistent and reliable indicator of understanding the fairness metrics. The comprehension score demonstrated in this work lays a foundation for many future studies exploring other fairness definitions.

We do find, however, that comprehension is lower for equal opportunity, false negative rate than other definitions. In general, comprehension scores for equal opportunity (both FNR and FPR) were less internally consistent than other fairness rules, suggesting participant responses were also more "noisy" for equal opportunity. This is somewhat intuitive: equal opportunity is difficult to understand, as it only involves one type of error (FNR or FPR) rather than both. Furthermore, FNR participants had the lowest comprehension scores *and* the lowest consistency of all conditions. We believe this finding also matches intuition: FNR is a strange notion in the context of hiring, as it concerns only those qualified applicants who were *not* hired or offered jobs. Indeed, in free-response questions several participants mentioned that they do not understand why qualified candidates are *not* hired. We believe many participants fixated on this strange setting, impacting their comprehension scores. This finding is potentially problematic, as equal opportunity definitions are increasingly used in practice. Indeed, major fairness tools such as Google What-If tool (Wexler et al., 2019) and the IBM AI Fairness 360 (Bellamy et al., 2019) specifically focus on equal opportunity. Further work should be put into making descriptions of nuanced fairness metrics more accessible.

Our analysis also identified other issues that should be considered when thinking about mathematical notions of fairness. First, we find that education is a strong predictor of comprehension. This is especially troubling, as the negative impacts of biased ML are expected to disproportionately impact the most marginalized (Barocas & Selbst, 2016) and displace employment opportunities for those with the least education (Frey & Osborne, 2017). Lack of understanding may hamper these groups' ability to effectively advocate for themselves. Designing more accessible explanations of fairness should be a top research priority.

Second, we find that those with the weakest comprehension of fairness metrics also express the least negative sentiment toward them. When fairness is a concern, there are always trade-offs—between accuracy and equity, or between different stakeholders, and so on. Balancing these trade-offs is an uncomfortable dilemma often lacking an objectively correct solution. It is possible that those who comprehend this dilemma *also* recognize the precarious trade-off struck by any mathematical definition of fairness, and are therefore dissatisfied with it. From another perspective, this finding is more insidious. If those with the weakest understanding of AI bias are also least likely to protest, then major problems in algorithmic fairness may remain uncorrected.

Acknowledgments

Dickerson, McElfresh, and Schumann were supported in part by NSF CAREER Award IIS-1846237, DARPA GARD Award #HR112020007, DARPA SI3-CMD Award #S4761, DoD WHS Award #HQ003420F0035, NIH R01 Award NLM-013039-01, and a Google Faculty Research Award. We gratefully acknowledge funding support from the NSF (Grants 1844462 and 1844518). The opinions in this paper are those of the authors and do not necessarily reflect the opinions of any funding sponsor or the United States Government.

References

Akaike, H. A new look at the statistical model identification. In *IEEE Transactions on Automatic Control*, volume 19, pp. 716–723, 1974. doi: 10.1109/TAC.1974.1100705.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. *ProPublica, May*, 23:2016, 2016.

Barocas, S. and Selbst, A. D. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Batanero, C., Chernoff, E. J., Engel, J., Lee, H. S., and Sánchez, E. Research on teaching and learning probability. In *Research on teaching and learning probability*, pp. 1–33. Springer, Cham, 2016.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

Binns, R. Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research*, 81:1–11, 2017.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., and Shadbolt, N. 'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 377. ACM, 2018.

Bureau, U. C. 2017 us census demographics, 2017. URL `https://data.census.gov/cedsci`.

Caliskan, A., Bryson, J. J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Chouldechova, A. and Roth, A. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

Cint. Cint. URL `https://www.cint.com/`.

Datta, A., Tschantz, M. C., and Datta, A. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

Everitt, B. and Skrondal, A. *The Cambridge Dictionary of Statistics*. Cambridge University Press, 4th edition, 2010.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268. ACM, 2015.

Frey, C. B. and Osborne, M. A. The future of employment: How susceptible are jobs to computerisation? *Technological forecasting and social change*, 114:254–280, 2017.

Gigerenzer, G. and Edwards, A. Simple tools for understanding risks: from innumeracy to insight. *Bmj*, 327 (7417):741–744, 2003.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., and Woloshin, S. Helping doctors and patients make sense of health statistics. *Psychological science in the public interest*, 8(2):53–96, 2007.

Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., and Weller, A. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*, pp. 903–912. International World Wide Web Conferences Steering Committee, 2018.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *NeurIPS*, pp. 3315–3323, 2016.

Harrell, M. C. and Bradley, M. A. Data collection methods. semi-structured interviews and focus groups. Technical report, Rand National Defense Research Inst santa monica ca, 2009.

Harrison, G., Hanson, J., Jacinto, C., Ramirez, J., and Ur, B. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 392–402, 2020.

Hogarth, R. M. and Soyer, E. Providing information for decision making: Contrasting description and simulation. *Journal of Applied Research in Memory and Cognition*, 4 (3):221–228, 2015.

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., and Baesens, B. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decis. Support Syst.*, 51(1):141–154, April 2011. ISSN 0167-9236. doi: 10.1016/j.dss.2010.12.003. URL http://dx.doi.org/10.1016/j.dss.2010.12.003.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

Lee, M. K. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1): 2053951718756684, 2018.

Lee, M. K. and Baykal, S. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pp. 1035–1048, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4335-0. doi: 10.1145/2998181.2998230. URL http://doi.acm.org/10.1145/2998181.2998230.

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., and Kusbit, D. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. In *Proc. ACM Hum.-Comput. Interact., 3, CSCW*, pp. Article 182, New York, NY, USA, 2019. ACM. URL https://doi.org/10.1145/3359284.

Lipton, Z. C. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43, 2018.

Nunnally, J. *Psychometric Theory*. McGraw-Hill, 2nd edition, 1978.

Plane, A. C., Redmiles, E. M., Mazurek, M. L., and Tschantz, M. C. Exploring user perceptions of discrimination in online targeted advertising. In *26th USENIX Security Symposium (USENIX Security 17)*, pp. 935–951, 2017.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5680–5689, 2017.

Rawls, J. *A Theory of Justice*. Harvard University Press, 1971.

Redmiles, E. M., Kross, S., and Mazurek, M. L. How well do my results generalize? comparing security and privacy survey results from mturk, web, and telephone samples. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 1326–1343. IEEE, 2019.

Ribeiro, M. T., Singh, S., and Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. ACM, 2016.

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., and Liu, Y. How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations. *Artificial Intelligence*, 283:103238, 2020.

Srivastava, M., Heidari, H., and Krause, A. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. *CoRR*, abs/1902.04783, 2019. URL http://arxiv.org/abs/1902.04783.

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., and Wilson, J. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.

Woodruff, A., Fox, S. E., Rousso-Schindler, S., and Warshaw, J. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 656. ACM, 2018.