Human Comprehension of Fairness in Machine Learning

Debjani Saha
University of Maryland
College Park, MD, USA
dsaha@cs.umd.edu

Candice Schumann
University of Maryland
College Park, MD, USA
schumann@cs.umd.edu

Duncan C. McElfresh
University of Maryland
College Park, MD, USA
dmcelfre@math.umd.edu

John P. Dickerson
University of Maryland
College Park, MD, USA
john@cs.umd.edu

Michelle L. Mazurek
University of Maryland
College Park, MD, USA
mmazurek@cs.umd.edu

Michael Carl Tschantz
ICSI
Berkeley, CA, USA
mct@icsi.berkeley.edu

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 non-technical audience understand a basic definition of ML fairness?* We develop a metric to measure comprehension of one such definition–demographic parity. We validate this metric using online surveys, and study the relationship between comprehension and sentiment, demographics, and the application at hand.

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, Transparancy, 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 [5, 8, 11]. 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, one major question underlying the study of ML fairness remains unanswered in the literature: *Does the general public understand these new, mathematical definitions of ML fairness along with their behavior in ML applications?*

Our Contributions. We take a first step to answering the above question by studying peoples' comprehension and perceptions of one popular definition of ML fairness, namely *demographic parity*. Specifically, we address the following research questions:

- **RQ1** Does a non-technical audience comprehend the definition and implications of demographic parity?
- **RQ2** Do demographics play a role in comprehension?
- **RQ3** How are comprehension and sentiment related?
- **RQ4** Does the application scenario affect comprehension?

We developed an online survey to address these four research questions. The survey was tailored to a non-technical audience. We present participants with one of three simple, but realistic, decision-making scenarios where fairness plays a role. Each scenario is accompanied by a *fairness rule* (corresponding to demographic parity), expressed in each scenario's context. We ask several 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.

Using two forms of validation, we find that this comprehension score is a consistent and reliable indicator of understanding demographic parity. Exploratory analysis reveals that education level is an important predictor for comprehension, and that *negative* sentiment is associated with *greater* comprehension of demographic parity. These findings inspire several areas for future work.

2 RELATED WORK

In response to many instances of bias in fielded artificial intelligence (AI)/machine learning (ML) systems, ML fairness has received significant attention from the computer science community. Notable examples include gender bias in job-related ads [9], racial bias in evaluating names on resumes [7], and racial bias in predicting criminal recidivism [2]. To correct this biased behavior, ML researchers have proposed several mathematical notions of fairness. We discuss some of these notions here, along with public understanding and perceptions of AI/ML decision-makers.

AI/ML Fairness. *Demographic parity* requires that the probability of some outcome $\hat{Y} = 1$ should be the same regardless of sensitive

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

group membership. In this exploratory study we focus on describing demographic parity to participants. We choose this particular fairness definition because it is popular in use and conceptually the simplest, making it an ideal candidate for developing a survey to measure comprehension of ML fairness and gathering initial insights in this field.

Like all proposed mathematical notions of fairness, demographic parity is not perfect.For example, Hardt et al. [16] criticize it while providing two new properties, equal opportunity and equalized odds. We believe that our survey methods can be adapted to these properties as well as other notions of fairness, such as calibration [26] and causal fairness [19].

The aforementioned prior work is written primarily by, and for, ML researchers. Some researchers (us included) have studied how the general public perceives AI/ML systems.

Perceptions of AI/ML Decision-Makers. Our work is most similar to a collection of studies on how people perceive AI/ML decision-makers. Lee [20] studies perceptions of fairness, trust, and emotional response of algorithmic decision-makers — as compared to human decision-makers. With others, she studied perceptions of fairness while splitting goods or tasks [21, 22]. Binns et al. [6] studies how different explanation styles impact perceptions of algorithmic decision-makers.

A related body of work focuses on *bias* in algorithmic systems. Woodruff et al. [28] investigates perceptions of algorithmic bias among marginalized populations, using a focus group-style workshop. Grgic-Hlaca et al. [15] studies the underlying factors causing perceptions of bias, highlighting the importance of selecting appropriate features in algorithmic decision-making. Plane et al. [25] look at perceptions of discrimination of online advertising.

Perceptions of algorithmic decision-making are closely related to peoples' *understanding* of the process. The related field of interpretable ML focuses on communicating the decision-making process and results of ML-based decisions [23]. Many tools have been developed to make ML models more interpretable, and many demonstrably improve understanding of ML-based decisions [18, 27].

Teaching probability. Since fairness metrics (including demographic parity) are often defined in terms of probability, while developing our survey we considered best practices for teaching and communicating probability concepts. Batanero et al. [4] provide an overview of teaching probability and how students learn probability. We follow their definition of a classical probability which relies on proportions. Gigerenzer and Edwards [13] find that numerical representations of probabilities can be confusing for humans. Gigerenzer et al. [14] observe that many physicians do not understand health statistics and probabilities. Hogarth and Soyer [17] suggest using simulated experiences of decisions to help understand probabilistic notions. Our decision to present fairness definitions in the context of simple, realistic scenarios (described in detail in §3) reflects this suggestion.

3 METHODS

To study perceptions of ML fairness, we conducted an online survey where participants are presented with a hypothetical decision-making scenario. The participants are then presented with a "rule" for enforcing fairness. We then ask each participant several questions on their comprehension and perceptions of this fairness rule.

We use three different decision-making scenarios – **Art Project (AP):** distributing awards for art projects amongst primary school students, **Employee Awards (EA):** distributing employee awards at a sales company, and **Hiring (HR):** distributing job offers to applicants. In each scenario the students/employees/applicants are partitioned into two groups (parents' occupation for the first scenario, and binary gender for the other two scenarios). The fairness rule is modeled after the notion of *demographic parity*. In short, this rule requires that the fraction of one group who receives a *positive* outcome (i.e., an award or job offer) is equal for both groups. We describe our survey design in greater detail in §3.1. This study was approved by our organization's standard ethical review process.

3.1 Survey Design

Here we provide a high-level discussion of the survey design; the full text of each survey can be found in Appendix A. We use a between-subjects design: participants are randomly partitioned into three groups, one for each scenario (AP, EA, or HR). The participant is first presented with a consent form (see Appendix B). If consent is obtained, the participant moves on to the survey, where a short paragraph explains the decision-making scenario. To make the notion of *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. For example, in HR we define the award 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.*

After presenting the scenario description and fairness rule we ask a series of 17 questions, including: 2 questions concerning participant evaluation of the scenario, 9 comprehension questions about the fairness rule, 2 self-report questions on participant understanding and use of the rule, and 4 free-response questions on both comprehension and sentiment. At the end of the survey, we collect some demographic information from participants (age, gender, race/ethnicity, education level, and expertise in a number of relevant fields).

3.2 Cognitive Interviews

We conducted in-person cognitive interviews to pilot our survey. We recruited 9 participants from a large metropolitan area using Craigslist. We required participants to be over 18 years of age and fluent in English. Participants ranged between the ages of 20 and 66. These interviews took place on our organization's campus and lasted about 1 hour. All participants signed a written consent form prior to the interview, and were paid $30 for their time.

During these interviews, participants completed a preliminary version of our survey (described in §3.1). After each survey question, we asked the participants several interview questions related to their comprehension of and feelings toward the survey. A primary finding of these interviews is that some participants tended to use their own personal notions of fairness when answering comprehension questions rather than using the definition we provided. We

	Census %	Obtained %
Ethnicity		
AI or AN	0.7	0.7
Asian or NH or PI	5.7	1.4
Black or AA	12.3	10.2
Hispanic or Latinx	18.1	12.2
Other	2.6	2.7
White	60.6	72.8
Education Level		
Less than HS	12.1	6.1
HS or equivalent	27.7	29.9
Post-secondary, no Bachelor's	30.8	30.6
Bachelor's and above	29.4	33.3

Table 1: Participant demographics across ethnicity and education level; target was to match the 2017 US Census. AI = American Indian, AN = Alaska Native, NH = Native Hawaiian, PI = Pacific Islander, AA = African American.

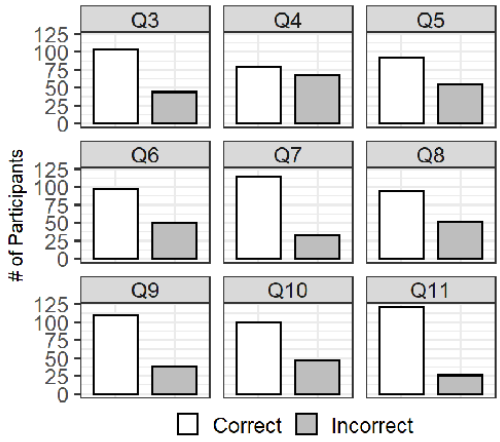


Figure 1: Number of participants answering each question correctly. Each panel contains all 147 participants.

were concerned that this would limit our ability to effectively measure comprehension. To address this problem, we rewrote several parts of our survey and added two new questions (Q14 and Q15).

3.3 Recruitment

We recruited participants using the online service Cint,[2] which allowed us to match the US 2017 Census distributions of ethnicity and education level to make sure that definitions used were not just evaluated by one demographic. We required that participants be 18 years of age or older, and fluent in English. Participants were compensated using Cint's rewards system, which is based on marketplace points. 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."

3.4 Participants

A total of 147 participants were included in the present analysis. This included 75 males (51.0%), 71 females (48.3%), and 1 (0.7%) preferring not to answer. The average age was 46 years old (SD = 16 years). The ethnicity and education level of our participants were intended to resemble that of the 2017 US Census.[3] See Table 1 for more details. On average, participants completed the online survey in 14 minutes.

3.5 Data Analysis

Free response questions were qualitatively coded for statistical testing. One question was coded by a single researcher for simple correctness (see §4.1). The other was triple-coded (resolved to 100%) to capture sentiment information (see §4.3).

The following methods were used for all statistical analyses unless otherwise specified. On nonparametric ordinal data, omnibus comparisons 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. Chi-squared tests were used for comparisons of nominal data. All boxplots show median and first and third quartiles; whiskers extend to $1.5 * \text{IQR}$ (interquartile range = Q3 − Q1), with outliers indicated by individual points.

[2] https://www.cint.com/
[3] https://data.census.gov/cedsci

4 RESULTS

We analyze survey responses and make several observations. As an important first step, we find that we can, in fact, measure comprehension of the fairness rule. Given this measurement of comprehension, we find that comprehension is potentially influenced by a participant's education level. We also find that participants exhibiting high comprehension tended to disagree with the rule, while those exhibiting low comprehension tended to not comply with rule application.

4.1 Our Survey Effectively Captures Rule Comprehension (RQ1)

We find that our survey (and resulting comprehension score) effectively measures participant comprehension of fairness definitions (i.e. rules). 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). Cronbach's α and item-total correlation were used to assess internal validity of these questions. Both measures met established thresholds for internal validity [10, 24]: Cronbach's $\alpha = 0.71$, and item-total correlation for all 9 items > 0.3, with the exception of one of the true/false questions (Q5, see Appendix A for full question). We may omit this question in future work (see §5). See Fig. 1 for participant performance per question.

The validity of our comprehension score as a measure of participant ability to understand a rule was interrogated using two self-report measures and one free response question.

4.1.1 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 (see Fig. 2). Using Spearman's Rho (appropriate for ordinal data), we find a correlation coefficient $\rho = 0.39$ ($p < 0.001$). This suggests that participants were accurately assessing their ability to apply the rule.

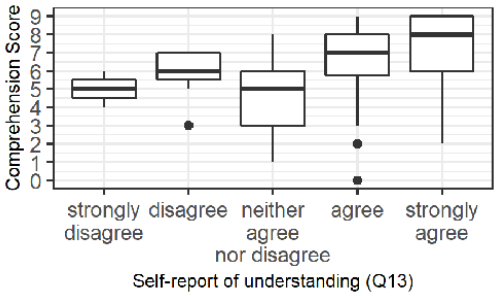


Figure 2: Comprehension score grouped by response to Q13. Self-reported understanding of the rule was associated with higher comprehension scores. X-axis is reversed for figure and correlation test.

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 added this question because during the cognitive interviews (see §3.2), several participants indicated that they were applying their own notion of fairness rather than following the provided rule. 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 = 7.09) than those who used their own notions (4.68) or a combination (4.90) thereof (post-hoc M-WU, $p < 0.001$ for both tests; 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. 3).

4.1.2 Participants with higher comprehension scores are better able to explain the rule. To further validate our comprehension score, 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.

The results of Q12 can be seen in Fig. 4. A K-W test revealed a relationship between comprehension score and coded responses to Q12 ($p < 0.001$). Both correct (mean comprehension score = 7.71) and partially correct (7.03) responses were associated with higher

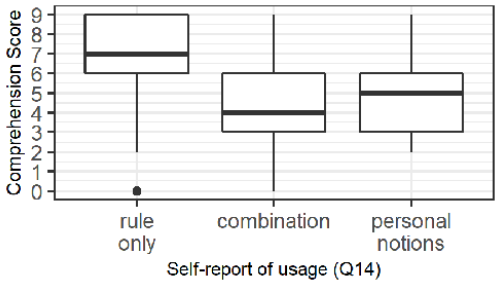


Figure 3: Comprehension score grouped by response to Q14. Rule compliance (leftmost on the x-axis) was associated with higher comprehension scores. One participant who did not provide a response was excluded from the figure and relevant analysis.

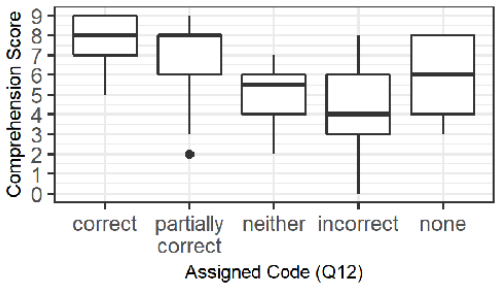


Figure 4: Comprehension score grouped by code assigned to Q12 response. Participants who provided either correct or partially correct responses tended to perform better.

comprehension scores than neither (5.13, $p < 0.001$ for both) and incorrect (4.24, $p < 0.001$ for both) responses (post-hoc M-WU; corrected $\alpha = 0.05/10 = 0.005$). There was no difference between correct and partially correct responses, nor between neither and incorrect responses. There was also no difference between none (5.80) and the other four codes.These findings further corroborate our claim that our comprehension score is a valid measure of fairness rule comprehension.

4.2 Education Influences Comprehension (RQ2)

During the cognitive interview phase, we observed a possible trend of comprehension scores being lower for older participants and those with less educational attainment. If true, this would suggest that fairness explanations should be carefully validated to ensure they can be used with diverse populations. We investigated this hypothesis, in an exploratory fashion, using poisson regression models.

Three models were tested. The first regressed score against all four demographic categories as predictors (gender, age, ethnicity, and education), the second omitted education, and the third tested only education. Models were compared using Akaike information criterion (AIC), a standard method of evaluating model quality and performing model selection [1]. Comparison by AIC revealed that model 1 (all four categories) was a better predictor for comprehension score than models 2 or 3 (AIC = 643.3, 651.2, and 660.5,

respectively; difference = 0.0, 7.9, and 17.1). In model 1, only education showed correlation with comprehension score (effect size = 1.40, $p < 0.05$). Further work is needed to confirm this exploratory result.

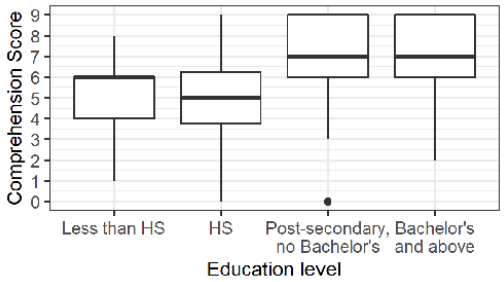


Figure 5: Comprehension score grouped by education level. Higher education level was associated with higher comprehension scores.

4.3 Disagreement with the Rule is Associated with Higher Comprehension Scores (RQ3)

Participants were asked for their opinion on the presented rule in another free response question (Q15). These responses were then qualitatively coded to capture participant sentiment towards the rule as 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 fairness of the rule.

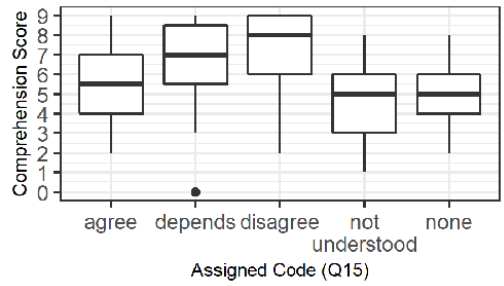


Figure 6: Comprehension score grouped by code assigned to Q15 response. Participants who exhibited negative sentiment toward the rule responses tended to perform better.

This question was added based on the cognitive interviews, where perception seemed to influence compliance. The results of Q15 can be seen in Fig. 6. A K-W test revealed a relationship between comprehension score and coded response to Q15 ($p < 0.001$). 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 all three comparisons; corrected $\alpha = 0.05/10 = 0.005$). No other differences were found.

Note that this result should not be interpreted as an overall finding on the appropriateness of demographic parity. Instead we anticipate the perceptions of appropriateness of any fairness definition will be highly context-dependent.

4.4 Non-Compliance is Associated with Lack of Understanding

We were interested in understanding the reason 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 the following analyses, conducted using χ^2 tests.

Non-compliant participants were less likely to self-report high understanding of the rule in Q13 ($p < 0.001$, see Fig. 7). Recall that both compliance and higher self-reported understanding were associated with higher comprehension score (see §4.1). Moreover, non-compliance appears to be associated with a reduced ability to correctly explain the rule in Q12 ($p < 0.001$, see Fig. 8). Recall also that correctness of the rule explanation was associated with higher comprehension score (see §4.1). Finally, negative participant sentiment towards the rule (Q15) appears to be associated with greater compliance ($p < 0.005$, see Fig. 9), as well as higher comprehension score (see §4.3). Thus, non-compliant participants appear to behave this way because they do not *understand* the rule, rather than because they do not *like* it.

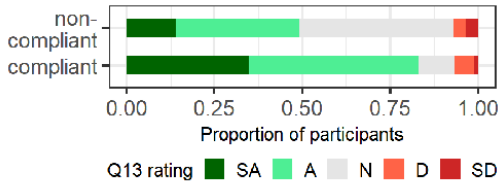


Figure 7: Self-report of understanding (Q13) split by compliance (Q14). NC participants tend to report less confidence in their ability to apply the rule. SA = strongly agree, A = agree, N = neither agree nor disagree, D = disagree, SD = strongly disagree.

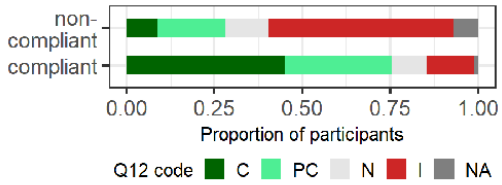


Figure 8: Correctness of rule explanation (Q12) split by compliance (Q14). NC participants tend to be less able to explain the presented rule in their own words. C = correct, PC = partially correct, N = neither, I = incorrect, NA = none.

4.5 Scenario does not Influence Comprehension Scores (RQ4)

One of our concerns going into this study was whether or not the nature of the scenario influences participants' ability to understand

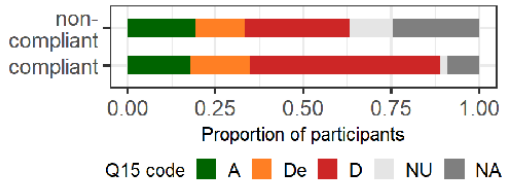


Figure 9: Participant agreement with rule (Q15) split by compliance (Q14). NC participants tend to harbor less negative sentiment towards the rule. A = agree, De = depends, D = disagree, NU = not understood, NA = none.

the fairness rule. Specifically, we were concerned that less realistic and/or important scenarios would cause participants to take the survey less seriously, and therefore perform more poorly. To test this, we devised three different surveys asking the same set of questions of three different scenarios (i.e., AP, EA, and HR; see §3 and Appendix A.1 for more detailed descriptions). Participants were randomly assigned to a scenario, resulting in the following distribution: AP = 41, EA = 49, HR = 57.

A K-W test revealed no differences between scenarios in terms of comprehension score (mean comprehension scores: AP = 6.0, EA = 6.74, HR = 5.86). However, differences did exist between scenarios in terms of importance (assessed in Q2), measured in hours of effort deemed necessary to make the relevant decision (K-W, $p < 0.001$). Post-hoc M-WU revealed that participants believed making a decision in the AP scenario merited fewer hours of effort (mean = 3.15hrs) than in the EA (13.52hrs, $p < 0.001$) or HR (15.23hrs, $p < 0.001$) scenarios (corrected $\alpha = 0.05/3 = 0.017$). See Fig. 10 for distributions of responses.

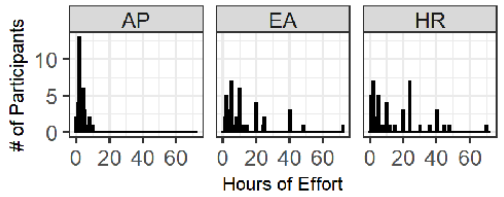


Figure 10: Importance of a scenario by proxy of hours of effort necessary to make a decision in each scenario. AP merited less hours of effort than both EA and HR.

Of note, it is possible that perceived realism, assessed in Q1 on a five-point Likert scale, was also influenced by scenario (K-W, $p = 0.051$), but we may need larger sample sizes to confirm this. Regardless, while the nature of a scenario does influence participant perception in terms of importance and (possibly) realism, it does not appear to influence comprehension (at least for the scenarios we chose).

5 DISCUSSION AND FUTURE WORK

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 one particular notion of ML fairness (demographic parity). We find that our comprehension score appears to be a consistent and reliable indicator of understanding demographic parity.

The comprehension score demonstrated in this work lays a foundation for many future studies. Perhaps the most obvious direction is to apply this comprehension score to other definitions of ML fairness, such as equal opportunity or calibration; indeed, this is a focus of our ongoing work.

In addition, our exploratory analysis raised several hypotheses that should be investigated in future work. First, we find that education is a major predictor of comprehension. We studied one of the simplest ML fairness definitions — demographic parity — yet there is a wide range of comprehension. This disparity will likely only worsen with more complicated definitions of fairness. This is especially troubling, as the negative impacts of biased ML is expected to disproportionally impact the most marginalized among us [3], and displace employment opportunities for those with the least education [12]. Designing more accessible explanations of fairness should be a top research priority.

Second, we find that those with the strongest comprehension of demographic parity also express the most negative sentiment toward it. We hypothesize that this effect would occur with other definitions of fairness, for an intuitive reason: 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 fully comprehend the dilemma at hand *also* comprehend the precarious trade-off struck by a 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

This research was supported by NSF IIS #1844462, NSF IIS #1844518, NSF IIS RI CAREER Award #1846237, and a generous gift from Google.

REFERENCES

[1] H Akaike. 1974. A new look at the statistical model identification. In *IEEE Transactions on Automatic Control*, Vol. 19. 716âĂŞ723. https://doi.org/10.1109/TAC.1974.1100705
[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016), 2016.
[3] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.
[4] Carmen Batanero, Egan J Chernoff, Joachim Engel, Hollylynne S Lee, and Ernesto Sánchez. 2016. Research on teaching and learning probability. In *Research on teaching and learning probability*. Springer, Cham, 1–33.
[5] Reuben Binns. 2017. Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of Machine Learning Research* 81 (2017), 1–11.

[6] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. '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*. ACM, 377.
[7] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.
[8] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
[9] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (2015), 92–112.
[10] BS Everitt and A Skrondal. 2010. *The Cambridge Dictionary of Statistics* (4th ed.). Cambridge University Press.
[11] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268.
[12] Carl Benedikt Frey and Michael A Osborne. 2017. The future of employment: How susceptible are jobs to computerisation? *Technological forecasting and social change* 114 (2017), 254–280.
[13] Gerd Gigerenzer and Adrian Edwards. 2003. Simple tools for understanding risks: from innumeracy to insight. *Bmj* 327, 7417 (2003), 741–744.
[14] Gerd Gigerenzer, Wolfgang Gaissmaier, Elke Kurz-Milcke, Lisa M Schwartz, and Steven Woloshin. 2007. Helping doctors and patients make sense of health statistics. *Psychological science in the public interest* 8, 2 (2007), 53–96.
[15] Nina Grgic-Hlaca, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 903–912.
[16] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NeurIPS*. 3315–3323.
[17] Robin M Hogarth and Emre Soyer. 2015. Providing information for decision making: Contrasting description and simulation. *Journal of Applied Research in Memory and Cognition* 4, 3 (2015), 221–228.
[18] Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. 2011. An Empirical Evaluation of the Comprehensibility of Decision Table, Tree and Rule Based Predictive Models. *Decis. Support Syst.* 51, 1 (April 2011), 141–154. https://doi.org/10.1016/j.dss.2010.12.003
[19] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.
[20] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.
[21] Min Kyung Lee and Su Baykal. 2017. 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)*. ACM, New York, NY, USA, 1035–1048. https://doi.org/10.1145/2998181.2998230
[22] Min Kyung Lee, Anuraag Jain, Hae Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. In *Proc. ACM Hum.-Comput. Interact., 3, CSCW*. ACM, New York, NY, USA, Article 182. https://doi.org/10.1145/3359284
[23] Zachary C Lipton. 2018. The mythos of model interpretability. *Commun. ACM* 61, 10 (2018), 36–43.
[24] JC Nunnally. 1978. *Psychometric Theory* (2nd ed.). McGraw-Hill.
[25] Angelisa C Plane, Elissa M Redmiles, Michelle L Mazurek, and Michael Carl Tschantz. 2017. Exploring user perceptions of discrimination in online targeted advertising. In *26th USENIX Security Symposium (USENIX Security 17)*. 935–951.
[26] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5680–5689.
[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. 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*. ACM, 1135–1144.
[28] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 656.

A SURVEY

Each of the surveys are split into four main sections. The first section is the consent form which can be found in Appendix B. The second section describes the scenario and asks questions about the given scenario (§A.1). The third section describes the fairness metric, defined as the rule, used (in this case it is demographic parity) and asks specific questions about the metric (§A.2). Finally the last section asks for demographic information (§A.3).

A.1 Scenario description and questions

We introduce three different decision making scenarios. Words that vary across scenario in the questions are shown as <art project, employee awards, hiring>.

A.1.1 Art project. A fourth grade teacher is reviewing 20 student art projects. They will award lollipops to the top 4 students who put the most effort into their projects. The teacher knows that some of the students have artists as parents, who might have helped their children with their art project. The teacherâĂŹs goal is to give out lollipops only based on the amount of effort that the student *themselves* put into their projects.

The teacher uses the following criteria to decide who should get a lollipop:

- Elaborateness of each project.
- Creativity of each project.

About 50% of the students have artists as parents, and 50% do not.

In the past, students with artists as parents typically put more effort into their projects.

In this group of students there is a wide range of project quality (as measured by elaborateness and creativity). However, this range of quality is about the same between students with artists as parents and those without.

The teacher wants to make sure that they award lollipops in a fair way, no matter whether the studentsâĂŹ parents are artists or not.

A.1.2 Employee awards. A manager at a sales company is deciding which of their 100 employees should receive each of 10 mid-year awards. The managerâĂŹs goal is to give awards to employees who *will* have high net sales at the end of the year.

The manager uses the following criteria to decide who should get an award:

- Recent performance reviews
- Mid-year net sales
- Number of years on the job

About 50% of the employees are men, and 50% are women.

In the past, men have achieved higher end-of-year net sales than women.

In this group of employees, there is a wide range of qualifications (as measured by performance reviews, mid-year net sales, and number of years on the job). However, this range of qualifications is about the same between male and female employees.

The manager wants to make sure that this awards process is fair to the employees, no matter their gender.

A.1.3 Hiring. A hiring manager at a new sales company is reviewing 100 new job applications. Each applicant has submitted a resume, and has had an interview. The manager will send job

offers to 10 out of the 100 applicants. Their goal is to make offers to applicants who will have high net sales after a year on the job.

The manager will use the following to decide which applicants should receive job offers:

- Interview scores
- Quality of recommendation letters
- Number of years of prior experience in the field

About 50% of the applicants are men, and 50% are women.

In the past, men have achieved higher net sales than women, after one year on the job.

In this applicant pool there is a wide range of applicant quality (as measured by interview scores, recommendation letters, and years of prior experience in the field). However, the range of quality is about the same for both male and female applicants.

The hiring manager wants to make sure that this hiring process is fair to applicants, no matter their gender.

A.1.4 Questions.

(1) To what extent do you agree with the following statement: a scenario similar to the one described above might occur in real life. [5 point Likert scale]

(2) How much effort should the <teacher, manager, hiring manager> put in to make sure this decision is fair? [short answer - number of hours]

A.2 Rule description and questions

Unless otherwise noted the rule description is shown above each of the questions for reference. Correct answers are noted in red.

A.2.1 Art project. The teacher uses the following award rule to distribute lollipops: *The fraction of students who receive lollipops that have artist parents should equal the fraction of students in the class that have artist parents. Similarly, the fraction of students who receive lollipops that do not have artist parents should equal the fraction of students in the class that do not have artist parents.*

Example 1: If 10 out of the 20 students in the class have artist parents, then 2 out of the 4 lollipops would be awarded to students with artist parents (and the remaining 2 would be awarded to students without artist parents).

Example 2: If 5 out of the 20 students in the class have artist parents, then 1 out of the 4 lollipops would be awarded to students with artist parents (and the remaining 3 would be awarded to students without artist parents).

In the next section, we will ask you some questions about the information you have just read. Please note that this is not a test of your abilities. We want to measure the quality of the description you read, not your ability to take tests or answer questions.

Please note that we ask you to apply and use ONLY the above award rule when answering the following questions. You will have an opportunity to state your opinions and feelings on the rule later in the survey.

(3) Suppose a different teacher is considering awarding lollipops to the whole 4th grade. There are 100 students with artist parents, and 200 students without artist parents. The teacher decides to award 10 lollipops to students with artist parents. **Assuming the teacher is required to use the award rule above**, how many students without artist parents need to receive lollipops?
   (a) 10
   (b) 20
   (c) 40
   (d) 50

(4) **Assuming the teacher is required to use the award rule above**, in which of these cases can a teacher award more lollipops to students without artist parents than to students with artist parents?
   (a) When the students without artist parents have higher-quality projects (i.e., more elaborate and more creative) than those with artist parents.
   (b) When there are more students without artist parents than those with artist parents.
   (c) When students without artist parents have more creative projects than those with artist parents.
   (d) This cannot happen under the award rule.

(5) **Assuming the teacher is required to use the award rule above**, is the following statement TRUE OR FALSE: Even if a student with artist parents has a project that is of the same quality (i.e., equally elaborate and equally creative) as another project by a student without artist parents, they can be treated differently (ie., only one of the students might get a lollipop).

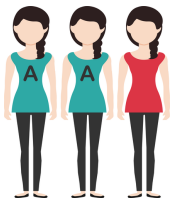
(6) **Assuming the teacher is required to use the award rule above**, is the following statement TRUE OR FALSE: If all students without artist parents have low-quality projects (i.e., low elaborateness and low creativity), but the teacher awards lollipops to some of them, then any lollipops awarded to students with artist parents must be awarded to those who have low-quality projects.

(7) **Assuming the teacher is required to use the award rule above**, is the following statement TRUE OR FALSE: Suppose the teacher is distributing 10 lollipops amongst a pool of students that includes students with and without artist parents. Even if all students with artist parents have low-quality (i.e., low elaborateness and low creativity) projects, some of them must still receive lollipops.

(8) **Assuming the teacher is required to use the award rule above**, is the following statement TRUE OR FALSE: This award rule always allows the teacher to award lollipops exclusively to the students who have the highest quality (i.e., most elaborate and most creative) projects.

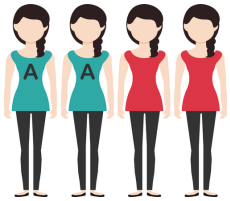
In the two examples above there are 20 students. Consider a different scenario, with **6 students – 4 with artist parents and 2 without, as illustrated below**. The next three questions each give a potential outcome for all six students (i.e., which of the 6 students receive awards). Please indicate which of the outcomes follow **the award rule above**.

(9) Alternative scenario 1:



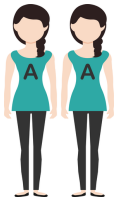
Does this distribution of awards obey the **award rule**? Yes

(10) Alternative scenario 2:



Does this distribution of awards obey the **award rule**? No

(11) Alternative scenario 3:



Does this distribution of awards obey the **award rule**? No

(12) In your own words, explain the **award rule**. [short answer] (The rule is not shown above this question)

(13) To what extent do you agree with the following statement: I am confident I know how to **apply the award rule described above**? [5 point Likert scale]

(14) Please select the choice that best describes your experience: When I answered the previous questions.
   (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.
   (c) I used a combination of the provided award rule and my own ideas of what the correct award decision should be.

(15) What is your opinion on the award rule? Please explain why. [short answer]

(16) Suppose that you are the teacher whose job it is to distribute lollipops to students based on the criteria listed above (i.e., elaborateness of each project, creativity of each project). How would you ensure that this process is fair? [short answer]

(17) Was there anything about this survey that was hard to understand or answer? [short answer]

A.2.2 Employee awards. The manager uses the following award rule to distribute awards: *The fraction of employees who receive awards that are female should equal the fraction of employees that are female. Similarly, fraction of employees who receive awards that are male should equal the fraction of employees that are male.*

Example 1: If there are 50 female employees out of 100, then 5 out of the 10 awards should be awarded to female employees (and the remaining 5 would be made to male employees).

Example 2: If there are 30 female employees out of 100, then 3 out of the 10 awards should be awarded to female employees (and the remaining 7 would be made to male employees).

In the next section, we will ask you some questions about the information you have just read. Please note that this is not a test of your abilities. We want to measure the quality of the description you read, not your ability to take tests or answer questions.

Please note that we ask you to apply and use ONLY the above award rule when answering the following questions. You will have an opportunity to state your opinions and feelings on the rule later in the survey.

(3) Suppose a different manager is considering employees for a different award. There are 100 male employees and 200 female employees, and they decide to give awards to 10 male employees. **Assuming the manager is required to use the award rule above**, how many female employees do they need to give awards to?
   (a) 10
   (b) 20
   (c) 40
   (d) 50

(4) **Assuming the manager is required to use the award rule above**, in which of these cases can a manager give more awards to female employees than to male employees?
   (a) When there are more well-qualified female employees than well-qualified male employees (i.e., more women have better performance reviews, higher mid-year net sales, and more years on the job).
   (b) When there are more female employees than male employees.
   (c) When female employees receive higher performance reviews than male employees.
   (d) This cannot happen under the award rule.

(5) **Assuming the manager is required to use the award rule above**, is the following statement TRUE OR FALSE: Even if a male employeeâĂŹs qualifications look similar to a female employeeâĂŹs (in terms of performance reviews, mid-year net sales, and years on the job), he can be treated differently (i.e., only one of the employees gets an award).

(6) **Assuming the manager is required to use the award rule above**, is the following statement TRUE OR FALSE: If all female employees are unqualified (i.e., have low performance reviews, low mid-year net sales, and few years on the job), but you give awards to some of them, then awards given to male employees must be made to unqualified male employees.

(7) **Assuming the manager is required to use the award rule above**, is the following statement TRUE OR FALSE: Suppose the manager is distributing 10 awards amongst a

pool that includes both male and female employees. Even if all male employees are unqualified for an award (i.e., have low performance reviews, low mid-year net sales, and few years on the job), some of them must still receive awards.

(8) **Assuming the manager is required to use the award rule above**, is the following statement TRUE OR FALSE: This award rule always allows the manager to distribute awards exclusively to the most qualified employees (i.e., employees with better performance reviews, high mid-year net sales, and high number of years on the job).

In the two examples above there are 100 employees. Consider a different scenario, with **6 employees– 4 female and 2 male, as illustrated below**. The next three questions each give a potential outcome for all six employees (i.e., which of the 6 employees receive awards). Please indicate which of the outcomes follow **the award rule above**.

(9) Alternative scenario 1:

Does this distribution of awards obey the **award rule**? Yes

(10) Alternative scenario 2:

Does this distribution of awards obey the **award rule**? No

(11) Alternative scenario 3:

Does this distribution of awards obey the **award rule**? No

(12) In your own words, explain the **award rule**. [short answer] (The rule is not shown above this question)

(13) To what extent do you agree with the following statement: I am confident I know how to **apply the award rule described above**? [5 point Likert scale]

(14) Please select the choice that best describes your experience: When I answered the previous questions.
   (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.
   (c) I used a combination of the provided award rule and my own ideas of what the correct award decision should be.

(15) What is your opinion on the award rule? Please explain why. [short answer]

(16) Suppose that you are the manager whose job it is to distribute mid-year awards to employees based on the criteria listed above (i.e., recent performance reviews, mid-year net sales, number of years on the job). How would you ensure that this process is fair? [short answer]

(17) Was there anything about this survey that was hard to understand or answer? [short answer]

A.2.3 Hiring. The hiring manager uses the following hiring rule to send out offers: *The fraction of applicants who receive job offers that are female should equal the fraction of applicants that are female. Similarly, fraction of applicants who receive job offers that are male should equal the fraction of applicants that are male.*

Example 1: If there are 50 female applicants out of the 100 applicants, then 5 out of the 10 offers would be made to female applicants (and the remaining 5 would be made to male applicants).

Example 2: If there are 30 female applicants out of the 100 applicants, then 3 out of the 10 offers would be made to female applicants (and the remaining 7 would be made to male applicants).

In the next section, we will ask you some questions about the information you have just read. Please note that this is not a test of your abilities. We want to measure the quality of the description you read, not your ability to take tests or answer questions.

Please note that we ask you to apply and use ONLY the above hiring rule when answering the following questions. You will have an opportunity to state your opinions and feelings on the rule later in the survey.

(3) Suppose a different hiring manager is considering applicants for a different job. There are 100 male applicants and 200 female applicants, and they decide to send offers to 10 male applicants. **Assuming the hiring manager is required to use the hiring rule above**, how many female applicants do they need to send offers to?
   (a) 10
   (b) 20
   (c) 40
   (d) 50

(4) **Assuming the hiring manager is required to use the hiring rule above**, in which of these cases can a hiring manager make more job offers to female applicants than to male applicants?
   (a) When there are more well-qualified female applicants than well-qualified male applicants (i.e., more women have higher interview scores, higher quality recommendation letters, and more years of prior experience in the field).

(b) When there are more female applicants than male applicants.
(c) When female applicants receive better interview scores than male applicants.
(d) This cannot happen under the hiring rule.

(5) **Assuming the hiring manager is required to use the hiring rule above**, is the following statement TRUE OR FALSE: Even if a male applicantâĂŹs qualifications look similar to a female applicantâĂŹs (in terms of interview scores, recommendation letters, and years of prior experience in the field), he can be treated differently (i.e., only one of the applicants will receive a job offer).

(6) **Assuming the hiring manager is required to use the hiring rule above**, is the following statement TRUE OR FALSE: If all female applicants are unqualified (i.e., have low interview scores, low-quality recommendation letters, and few years of prior experience in the field), but you send job offers to some of them, then any job offers made to male applicants must be made to unqualified male applicants.

(7) **Assuming the hiring manager is required to use the hiring rule above**, is the following statement TRUE OR FALSE: Suppose the hiring manager is sending out 10 job offers to a pool that includes male and female applicants. Even if all male applicants are unqualified (i.e., have low interview scores, low-quality recommendation letters, and few years of prior experience in the field), some of them must still receive job offers.

(8) **Assuming the hiring manager is required to use the hiring rule above**, is the following statement TRUE OR FALSE: This hiring rule always allows the hiring manager to send offers exclusively to the most qualified applicants (i.e., applicants with high interview scores, high quality recommendation letters, and high number years of prior experience in the field).

In the two examples above there are 100 applicants. Consider a different scenario, with **6 applicants – 4 female and 2 male, as illustrated below**. The next three questions each give a potential outcome for all 6 applicants (i.e., which of the 6 applicants receive job offers). Please indicate which of the outcomes follow **the hiring rule above**.

(9) Alternative scenario 1:

Does this distribution of job offers obey the **hiring rule**?
Yes

(10) Alternative scenario 2:

Does this distribution of job offers obey the **hiring rule**?
No

(11) Alternative scenario 3:

Does this distribution of job offers obey the **hiring rule**?
No

(12) In your own words, explain the **hiring rule**. [short answer] (The rule is not shown above this question)

(13) To what extent do you agree with the following statement: I am confident I know how to **apply the hiring rule described above**? [5 point Likert scale]

(14) Please select the choice that best describes your experience: When I answered the previous questions.
(a) I applied the provided hiring rule only.
(b) I used my own ideas of what the correct hiring decision should be rather than the provided hiring rule.
(c) I used a combination of the provided hiring rule and my own ideas of what the correct hiring decision should be.

(15) What is your opinion on the hiring rule? Please explain why. [short answer]

(16) Suppose that you are the hiring manager whose job it is to send job offers to applicants based on the criteria listed above (i.e., interview scores, quality of recommendation letters, number of years of prior experience in the field). How would you ensure that this process is fair? [short answer]

(17) Was there anything about this survey that was hard to understand or answer? [short answer]

A.3 Demographic Information

(1) Please specify the gender with which you most closely identify:
- Male
- Female
- Other
- Prefer not to answer

(2) Please specify your year of birth

(3) Please specify your ethnicity (you may select more than one):
- White
- Hispanic or Latinx

- Black or African American
- American Indian or Alaska Native
- Asian, Native Hawaiian, or Pacific Islander
- Other

(4) Please specify the highest degree or level of school you have completed:
- Some high school credit, no diploma or equivalent
- High school graduate, diploma or the equivalent (for example: GED)
- Some college credit, no degree
- Trade/technical/vocational training
- AssociateâĂŹs degree
- BachelorâĂŹs degree
- MasterâĂŹs degree
- Professional or doctoral degree (JD, MD, PhD)

(5) How much experience do you have in each of the following areas? (1 - no experience, 2 - limited experience, 3 - significant experience, 4 - expert)
- Human resources (making hiring decisions)
- Management (of employees)
- Education (teaching)
- IT infrastructure/systems administration
- Computer science/programming
- Machine learning/data science

We will maintain privacy of the information you have provided here. Your information will only be used for data analysis purposes.

B CONSENT

B.1 Online Survey Consent Form

B.1.1 Project Title. Fairness Evaluation and Comprehension

B.1.2 Purpose of the Study. This research is being conducted by [Blinded] at [Blinded]. We are inviting you to participate in this research project because you are above 18. The purpose of this research project is to understand lay comprehension of different fairness metrics.

B.1.3 Procedures. The procedures will start with reading a brief description of a decision-making scenario. You will then be asked to answer some comprehension questions about the scenario. The questions will look like the following: What are the pros and cons of the notion of fairness described above?

Finally, you will be asked some demographics questions. The entire survey will take approximately 20 minutes or less.

B.1.4 Potential Risks and Discomforts. There are several questions to answer over the course of this study, so you may find yourself growing tired towards the end. Outside of this, there are minimal risks to participating in this research study. All data collected in this study will be maintained securely (see Confidentiality section) and will be deleted at the conclusion of the study.

However, if at any time you feel that you wish to terminate your participation for any reason, you are permitted to do so.

B.1.5 Potential Benefits. There are no direct benefits from participating in this research. We hope that, in the future, other people might benefit from this study through improved understanding of fairness metrics and their applications.

B.1.6 Confidentiality. Any potential loss of confidentiality will be minimized by storing all data (including information such as MTurk IDs and demographics) will be stored securely (a) in a password-protected computer located at [Blinded] or (b) using a trusted third party (Qualtrics). Personally identifiable information that is collected (MTurk IDs, IP addresses, cookies) will be deleted upon study completion. All other data gathered will be stored for three years post study completion, after which it will be erased. The only persons that will have access to the data are the Principle Investigator and the Co-Investigators.

If we write a report or article about this research project, your identity will be protected to the maximum extent possible. Your information may be shared with representatives of the [Blinded] or governmental authorities if you or someone else is in danger or if we are required to do so by law.

B.1.7 Compensation. You will receive $3. You will be responsible for any taxes assessed on the compensation.

If you will earn $100 or more as a research participant in this study, you must provide your name, address and SSN to receive compensation.

If you do not earn over $100 only your name and address will be collected to receive compensation.

B.1.8 Right to Withdraw and Questions. Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you decide not to participate in this study or if you stop participating at any time, you will not be penalized or lose any benefits to which you otherwise qualify.

If you decide to stop taking part in the study, if you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the investigator: [Blinded]

B.1.9 Participant Rights. If you have questions about your rights as a research participant or wish to report a research-related injury, please contact:

[Blinded]

For more information regarding participant rights, please visit: [Blinded]

This research has been reviewed according to the [Blinded] IRB procedures for research involving human subjects.

B.1.10 Statement of Consent. By agreeing below you indicate that you are at least 18 years of age; you have read this consent form or have had it read to you; your questions have been answered to your satisfaction and you voluntarily agree to participate in this research study. Please ensure you have made a copy of the above consent form for your records.

Pease ensure you have made a copy of the above consent form for your records. A copy of this consent form can be found here [link to digital copy].

- I am age 18 or older
- I have read this consent form
- I voluntarily agree to participate in this research study

B.2 Cognitive Interview Consent Form

B.2.1 Project Title. Fairness Cognitive Interview

B.2.2 Purpose of the Study. This research is being conducted by [Blinded] at [Blinded]. We are inviting you to participate in this research project because you are above the age of 18, and fluent in English. The purpose of this research project is to understand lay comprehension of different fairness metrics.

B.2.3 Procedures. The procedure involves completing an interview. The full procedure will be approximately 1 hour in duration.

During the interview you will be audio recorded, if you agree to be recorded. You will be asked to first read a brief description of a decision-making scenario. You will then be asked to fill out a survey about the scenario. While answering questions you will be asked verbal questions related to how you reached your answer in the survey.

Sample survey question: Is the following statement true or false? This hiring rule allows the hiring manager to send offers exclusively to the most qualified applicants.

Sample interview question: How did you reach your answer to that survey question?

B.2.4 Potential Risks and Discomforts. There are several questions to answer over the course of this study, so you may find yourself growing tired towards the end. Outside of this, there are minimal risks to participating in this research study. All data collected in this study will be maintained securely (see Confidentiality section) and will be deleted at the conclusion of the study.

However, if at any time you feel that you wish to terminate your participation for any reason, you are permitted to do so.

B.2.5 Potential Benefits. There are no direct benefits from participating in this research. We hope that, in the future, other people might benefit from this study through improved understanding of fairness metrics and their applications.

B.2.6 Confidentiality. Any potential loss of confidentiality will be minimized by storing all data (including information such as demographics) securely (a) in a password protected computer located at [Blinded] or (b) using a trusted third party (Qualtrics). Personally identifiable information that is collected will be deleted upon study completion. All other data gathered will be stored for three years post study completion, after which it will be erased. The only persons that will have access to the data are the principle Investigator and the Co-Investigators.

If we write a report or article about this research project, your identity will be protected to the maximum extent possible. Your information may be shared with representatives of [Blinded] or governmental authorities if you or someone else is in danger or if we are required to do so by law.

B.2.7 Compensation. You will receive $30. You will be responsible for any taxes assessed on the compensation.

If you will earn $100 or more as a research participant in this study, you must provide your name, address and SSN to receive compensation.

If you do not earn over $100 only your name and address will be collected to receive compensation.

B.2.8 Right to Withdraw and Questions. Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you decide not to participate in this study or if you stop participating at any time, you will not be penalized or lose any benefits to which you otherwise qualify.

If you decide to stop taking part in the study, if you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the investigator: [Blinded]

B.2.9 Participant Rights. If you have questions about your rights as a research participant or wish to report a research-related injury, please contact: [Blinded]

For more information regarding participant rights, please visit: [Blinded]

This research has been reviewed according to [Blinded] IRB procedures for research involving human subjects.

B.2.10 Statement of Consent. Your signature indicates that you are at least 18 years of age; you have read this consent form or have had it read to you; your questions have been answered to your satisfaction and you voluntarily agree to participate in this research study. You will receive a copy of this signed consent form.

Please initial all that apply (you may choose any number of these statements):

- I agree to be audio recorded
- I agree to allow researchers to use my audio recording in research publications and presentations.
- I do not agree to be audio recorded

If you agree to participate, please sign your name below.