How Do Fairness Definitions Fare?
Examining Public Attitudes Towards Algorithmic Definitions of Fairness

Nripsuta Ani Saxena [1] Karen Huang [2] Evan DeFilippis [2] Goran Radanovic [2] David C. Parkes [2] Yang Liu [3]

Abstract

How should algorithmic fairness be defined? While many definitions of fairness have been proposed in the computer science literature, there is no clear agreement over one definition. We investigate ordinary people's perceptions of three of these fairness definitions. Across three online experiments, we test which definitions people perceive to be the fairest in the context of loan decisions, and whether fairness perceptions change with the addition of sensitive information (race or gender). One definition (calibrated fairness) tends to be preferred more, and the results also provide support for the principle of affirmative action.

1. Introduction

With the increasing pervasiveness of automated decision-making systems, there's a growing concern among computer scientists and the public about how to ensure algorithms are fair. While several definitions of fairness have recently been proposed in the computer science literature, there's a lack of agreement among researchers about which definition is the most appropriate (Gajane & Pechenizkiy, 2017). It is unlikely that one definition of fairness will be sufficient. This is supported also by recent impossibility results that show some fairness definitions cannot coexist (Kleinberg et al., 2016). Since the public is affected by these algorithmic systems, it is important to investigate public views of algorithmic fairness (Lee & Baykal, 2017; Lee et al., 2017; Lee, 2018; Binns et al., 2018; Woodruff et al., 2018).

While substantial research has been done in moral psychology to understand people's perceptions of fairness (e.g, Yaari & Bar-Hillel 1984, Bazerman et al. 1995), relatively little work has been done to understand how the general public views fairness criteria in algorithmic decision making: Pierson (2017) investigate how two different factors influence views on algorithmic fairness, and Binns et al. (2018) examine people's perception of justice in algorithmic decision making under different explanation styles. In contrast, our goal is to understand how people perceive the fairness definitions proposed in the recent computer science literature, that is, the outcomes allowed by these definitions. We look at fairness perceptions among the U.S. population.

1.1. Definitions of Fairness

Broadly, we investigate a concept of fairness known as *distributive justice*, or fairness regarding the outcomes (Adams, 1963; 1965). However, which characteristics regarding the individual should be relevant and which should be irrelevant to fairness? We investigate two characteristics: task-specific similarity (loan repayment rate) and a sensitive attribute (race or gender), and collect data on attitudes toward the relevancy of these characteristics. In principle, fairness is the absence of any bias based on an individual's inherent or acquired characteristics that are irrelevant in the context of decision-making (Chouldechova, 2017). In many contexts, these inherent characteristics (referred to as 'sensitive' or 'protected attributes' in the computer science literature), are gender, religion, race, skin color, age, and national origin.

We look at three fairness definitions from the computer science literature because they can be easily operationalized as decisions in the context of loans that are easily understood by lay people. We map these definitions (or constrained versions of them) to loan allocation choices, and test people's judgments of these choices. The definitions are:

Treating similar individuals similarly. Dwork et al. (2012) formulate fairness as treating similar individuals (with respect to certain attributes) similarly in receiving a decision. The similarity of any two individuals is determined on the basis of a similarity (distance) metric, specific to the task that ideally represents a notion of ground truth in regard to the decision context. An algorithm would be fair if its decisions satisfied the Lipschitz condition (a continuity and similarity measure) defined with respect to this given metric. In our experiments, individuals with similar repayment rates should receive similar amounts of money.

Never favor a worse individual over a better one. In

[1] University of Southern California, USA [2] Harvard University, USA [3] University of California, Santa Cruz, USA. Correspondence to: Nripsuta Ani Saxena <nsaxena@usc.edu>, Yang Liu <yangliu@ucsc.edu>.

Appearing at the International Conference on Machine Learning AI for Social Good Workshop, Long Beach, United States, 2019.

the context of online learning, Joseph et al. (2016) define fairness in a setting where one individual is to be selected for a favorable decision, as always choosing a better individual (with higher expected value of some measure of inherent quality) with a probability greater than or equal to the probability of choosing a worse individual. It promotes meritocracy with respect to an individual's inherent quality. In our experiments, an individual with a higher repayment rate should obtain at least as much money as her peer.

Calibrated fairness. Liu et al. (2017) formulate fairness in the setting of sequential decision-making. [1] This definition selects individuals in proportion to their merit. When the merit is known (underlying true quality), calibrated fairness implies the meritocratic fairness of Joseph et al. (2016). For a suitably chosen similarity metric, calibrated fairness implies Dwork et al. (2016). In our experiments, we interpret this definition as requiring that two individuals with repayment rates r_1 and r_2, respectively, should obtain $r_1/(r_1+r_2)$ and $r_2/(r_1+r_2)$ amount of money, respectively. [2]

For the purpose of the study, we need to interpret these fairness definitions, which are formalized for choosing a single individual for a favorable decision (or assigning an indivisible good) to this setting where the good is divisible.

2. Study 1 (No Sensitive Information)

In this study, our motivation is to investigate how information on an individual task-specific feature (i.e., the candidates' loan repayment rate) influences perceptions of fairness. We present participants with a scenario in which two individuals have each applied for a loan. The participants know no personal information about the two individuals except their loan repayment rates. We choose three allocation decisions, described below, that allow us to formulate qualitative judgments regarding the three fairness definitions.

2.1. Procedure

We want to understand how support for the fairness definitions depends on variation in the similarity of the target individuals. The three definitions differ in how this comparison between task-specific metrics should matter. Our experiments employed a between-subjects design with four conditions. We varied the individual candidates' similarity (dissimilarity) in ability to pay back their loan (i.e., their loan repayment rate), as an operationalization of task-specific similarity (dissimilarity). Participants were randomly shown one of four loan repayment rates: 55% and 50% (Treatment 1), 70% and 40% (Treatment 2), 90% and 10% (Treatment 3), and 100% and 20% (Treatment 4). We held all other information about the two candidates constant. Each participant was only shown one Treatment. We recruited 200 participants from Amazon Mechanical Turk (MTurk), and presented them with three possible decisions for how to allocate the money between the two candidates. The order of the decisions was randomized. Each decision was designed to help us to untangle the three fairness definitions.

"All A" Decision. Give all the money to the candidate with the higher payback rate. This decision is allowed in all treatments under meritocratic fairness as defined Joseph et al. (2016), where a worse applicant is never favored over a better one. It would also be allowed under the definition formulated by Dwork et al. (2012), in the more extreme treatments, and even in every treatment in the case that the similarity metric was very discerning. This decision would not be allowed in any treatment under the calibrated fairness definition (Liu et al., 2017).

"Equal" Decision. Split the money 50/50 between the candidates, giving $25,000 to each. This decision is allowed in all treatments under Dwork et al. (2012) – treating similar people similarly. Under this definition, when two individuals are deemed to be similar to each other, this is the *only* allowable decision. This decision is also allowed in all the treatments under the meritocratic definition (Joseph et al., 2016), as the candidate with the higher loan repayment rate is given at least as much as the other candidate, and hence is weakly favored. This decision would not be allowed in any treatment under calibrated fairness (Liu et al., 2017), since the candidates are not being treated in proportion of their quality (loan repayment rate).

"Ratio" Decision. Split the money between two candidates in proportion of their loan repayment rates. This decision is allowed in all treatments under calibrated fairness, where resources are divided in proportion to the true quality of the candidates. Moreover, this is the only decision allowed under this definition. This decision could also align with the definition proposed by Dwork et al. (2012), but only for suitably defined similarity metrics that allow the distance between decisions implied by the ratio allocation. Finally, this decision would be allowed under meritocratic fairness (Joseph et al., 2016) for the same reasons as the "Equal" decision. Namely, the candidate with the higher loan repayment rate is weakly favored to the other candidate.

We are testing human perceptions regarding the outcomes that different fairness definition allow, not the definitions themselves. However, if a certain definition allows multiple decisions, then we would expect these decisions to receive similar support. Where the perception of the fairness of out-

[1] Note that Kleinberg et al. (2016), Chouldechova (2017) define 'calibration' in a different way, that includes the notion of a sensitive attribute.

[2] This is a slightly different version of the formal definition in Liu et al. (2017), which would take the ratio in proportion to the rate at which one individual repays while the other does not, but we feel a more intuitive way to capture the idea of calibrated fairness in our setting.

comes is inconsistent with the allowable decisions for a rule, this is worthwhile to understand. If it is true that participants most prefer the treating similar people similarly definition, one would expect that they would prefer the “Equal” decision over the others for a wider range of similarity metrics and treatments. If it is true that participants most prefer the meritocratic definition, one would expect no significant difference in support for the three decisions. If it is true that participants most prefer the calibrated fairness definition, one would expect that the “Ratio” decision is perceived as more fair than the others.

We formulated the following set of hypotheses:

Hypothesis 1A. Across all treatments, participants perceive the “Ratio” decision as more fair than the “Equal” decision.

Hypothesis 1B. Across all treatments, participants perceive the “Ratio” decision as more fair than the “All A” decision.

Hypothesis 2. Participants perceive the “Equal” decision as more fair than the “All A” decision in Treatment 1. That is, participants may view the candidates in Treatment 1 as “similar enough” to be treated similarly.

Hypothesis 3. Participants perceive the “All A” decision as more fair than the “Equal” decision in Treatments 3 and 4.

2.2. Results and Discussion

We found partial support for H1A: participants perceived dividing the $50,000 between the two individuals in proportion of their loan repayment rates (the “Ratio” decision) as more fair than splitting the $50,000 equally (the “Equal” decision) in Treatments 2, 3, and 4. H1B was also partially supported: participants rated the “Ratio” decision as more fair than the “All A” decision in Treatments 1 and 2 (see Figure 1). We also found that participants rated the “All A” decision as more fair than the “Equal” decision in Treatment 3, but not in Treatment 4 (see Figure 1).

Data from Study 1 suggests participants perceived the “Ratio” decision (the only decision that aligns with calibrated fairness) to be more fair than the “Equal” decision (the only decision that is always aligned with the treating people similarly definition). One possible explanation is that calibrated fairness implies treating people similarly for a similarity metric (Liu et al., 2017) that is based on a notion of merit. In Treatments 1 and 2, participants rated the “Ratio” decision to be more fair than the “All A” decision. Note that the meritocratic definition is the only definition that always allows the “All A” decision. No significant difference was discovered for Treatments 3 and 4, where one candidate has a much higher repayment rate.

We found that participants rated the “Equal” decision as more fair than the “All A” definition in Treatment 1 (see Fig. 1), supporting H2. When the difference in the individuals’ loan repayment rates was small (5%), participants rated the decision to divide the money equally between the individuals as more fair than giving all the money to the individual with the higher loan repayment rate. It can be said that participants viewed individuals to be similar enough to be treated similarly only in Treatment 1.

3. Study 2 (Sensitive Information on Race)

In this study, our motivation is to investigate how the addition of sensitive information (race) to information on an individual task-specific feature (loan repayment rate) influences perceptions of fairness. We employed the same experimental paradigm and tested the same hypotheses as in Study 1. In addition to providing information on the individuals’ loan repayment rates, we also mention their race. We held the gender of the candidates constant (both were male), and randomized race (black or white). We recruited a separate sample of 1800 participants from MTurk, none of whom had taken part in Study 1. We found that participants viewed the “Ratio” decision as more fair than the “Equal”

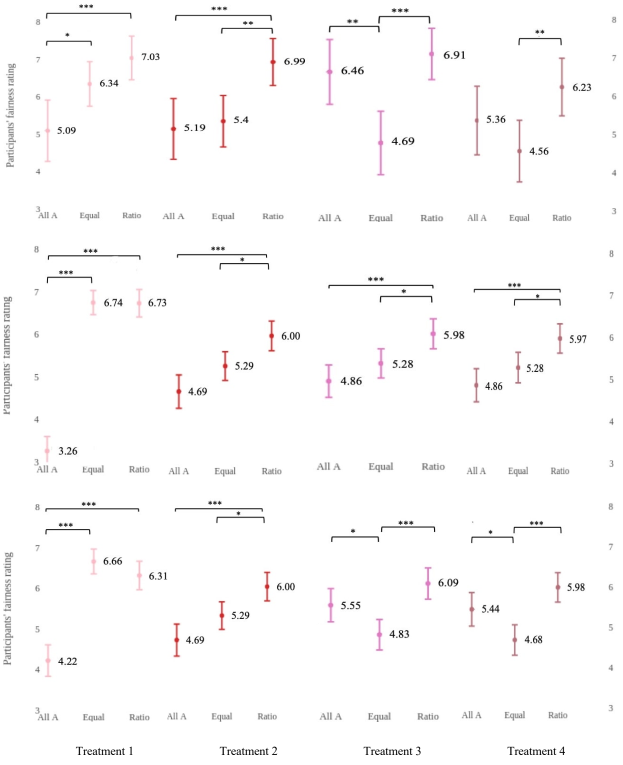

Figure 1. Comparison of means (with 95% CI). Top row: for Study 1; Middle: for Study 2 (when the individual with the higher loan repayment rate is white); Bottom row: for Study 2 (when the individual with the higher loan repayment rate is black). Where * signifies p <0.05, ** p <0.01, and *** p <0.001.

decision in Treatments 2, 3, and 4, regardless of race, in support of H1A. Further, we found an interaction effect for H1B: When the candidate with the higher repayment rate was white, people perceived the “Ratio” decision as more fair compared to the “All A” decision in all treatments. By contrast, when the candidate with the higher repayment rate was black, people perceived the “Ratio” decision as more fair compared to the “All A” decision only in Treatments 1 and 2 (See Fig. 1). Thus, participants in Study 2 gave most support to the decision to divide the $50,000 between the two individuals in proportion to their loan repayment rates, particularly when the individual with the higher loan repayment rate was white.

Further, participants rated the “Equal” decision as more fair than the “All A” decision in Treatment 1, regardless of race, in support of H2. This supports the corresponding results from Study 1, which indicate that one should account for similarity of individuals when designing fair rules. Importantly, we found evidence that race affects participants’ perceptions of fairness: participants showed the same preference (“Equal” more fair than “All A”) in Treatment 2, but only when the candidate with the higher repayment rate was white (see Fig. 1). We see further evidence of the effect of race: when the difference in loan repayment rates was larger (Treatments 3 and 4), participants rated the “All A” decision as more fair than the “Equal” decision, but only when the candidate with the higher repayment rate was black (see Fig. 1). These results suggest a boundary condition of H3: people may support giving all the loan money to the candidate with the higher payback rate, compared to splitting the money equally, when that candidate is a member of a group that is historically disadvantaged.

4. Study 3 (Sensitive Information on Gender)

In this study, we investigate if mentioning a different sensitive attribute (gender), instead of race, along with the candidates’ loan repayment rates influences perceptions of fairness. We employed the same experimental paradigm and tested the same hypotheses as in Study 2. We vary gender; race of the candidates was held constant (both were white), and randomized gender (male/female). A separate sample of 1800 participants was recruited from MTurk.

We found that participants viewed the “Ratio” decision as more fair than the “Equal” decision in Treatments 2, 3, and 4, regardless of gender, in support of H1A. Further, we observed an interaction effect for H1B: When the candidate with the higher repayment rate was male, people perceived the “Ratio” decision as more fair compared to the “All A” decision in all treatments. By contrast, when the candidate with the higher repayment rate was female, people perceived the “Ratio” decision as more fair compared to the “All A” decision only in Treatments 1, 2, and 3, but not in Treatment 4 (See Fig. 2). Overall, participants gave most support to the decision to divide the $50,000 between the two individuals in proportion to their loan repayment rates.

Further, participants rated the “Equal” decision as more fair than the “All A” decision in Treatment 1, regardless of gender, in support of H2 (see Fig. 2). This supports the corresponding results from Studies 1 and 2. Gender does have an effect: participants showed the same preference (“Equal” more fair than “All A”) in Treatment 2, but only when the candidate with the higher repayment rate was male. Furthermore, when the difference between the two candidates’ repayment rates was larger (Treatments 3 and 4), participants viewed the “All A” decision as more fair than the “Equal” decision, but only when the candidate with the higher repayment rate was female (see Fig. 2). This suggest the same boundary condition of H3 from Study 2: people show more support to giving all the loan money to the candidate with the higher repayment rate, compared to splitting the money equally, when that candidate is a member of a historically disadvantaged group.

5. Conclusion

People broadly show a preference for the “Ratio” decision, which is indicative of their support for the calibrated fairness definition (Liu et al., 2017) compared to the others. Race and gender do have an effect on people’s perceptions of fairness, and we also find some support for the principle of affirmative action. Understanding public attitudes is important for technologists and ethicists when designing algorithms that might affect the public.

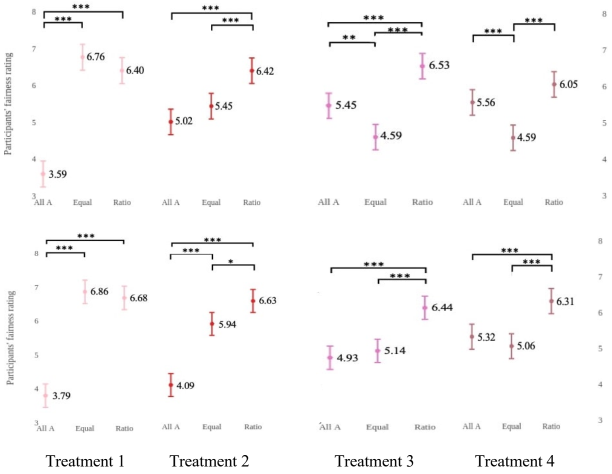

Figure 2. Comparison of means (with 95% CI). Top row: for Study 3 (when the individual with the higher loan repayment rate is male); Bottom: for Study 3 (when the individual with the higher loan repayment rate is female).

References

Adams, J. S. Towards an understanding of inequity. *The Journal of Abnormal and Social Psychology*, 67(5):422, 1963.

Adams, J. S. Inequity in social exchange. In *Advances in experimental social psychology*, volume 2, pp. 267–299. Elsevier, 1965.

Bazerman, M. H., White, S. B., and Loewenstein, G. F. Perceptions of fairness in interpersonal and individual choice situations. *Current Directions in Psychological Science*, 4(2):39–43, 1995.

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.

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

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. ACM, 2012.

Gajane, P. and Pechenizkiy, M. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017.

Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pp. 325–333, 2016.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

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 *CSCW*, pp. 1035–1048, 2017.

Lee, M. K., Kim, J. T., and Lizarondo, L. A human-centered approach to algorithmic services: Considerations for fair and motivating smart community service management that allocates donations to non-profit organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 3365–3376. ACM, 2017.

Liu, Y., Radanovic, G., Dimitrakakis, C., Mandal, D., and Parkes, D. C. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*, 2017.

Pierson, E. Gender differences in beliefs about algorithmic fairness. *arXiv preprint arXiv:1712.09124*, 2017.

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.

Yaari, M. E. and Bar-Hillel, M. On dividing justly. *Social choice and welfare*, 1(1):1–24, 1984.