Artificial Intelligence 283 (2020) 103238

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations[☆]

Nripsuta Ani Saxena[a,*], Karen Huang[b], Evan DeFilippis[b], Goran Radanovic[c,1], David C. Parkes[d], Yang Liu[e,*]

[a] Computer Science, University of Southern California, 941 Bloom Walk, Los Angeles, CA 90089, USA
[b] Organizational Behavior Unit, Harvard Business School, Soldiers Field Road, Boston, MA 02163, USA
[c] Computer Science, Max Planck Institute for Software Systems, Campus E1 5, D-66123 Saarbrücken, Germany
[d] Computer Science, Harvard University, Maxwell Dworkin, 33 Oxford St., Cambridge, MA 02138, USA
[e] Computer Science and Engineering, University of California, Santa Cruz, 1156 High St, Santa Cruz, CA 95064, USA

ARTICLE INFO

Article history:
Received 14 March 2019
Received in revised form 3 October 2019
Accepted 16 January 2020
Available online 20 February 2020

Keywords:
Fairness
Public attitudes
Human experiments
Algorithmic definition

ABSTRACT

What is the best way to define algorithmic fairness? While many definitions of fairness have been proposed in the computer science literature, there is no clear agreement over a particular definition. In this work, 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 (i.e., race or gender of the loan applicants). Overall, one definition (calibrated fairness) tends to be more preferred than the others, and the results also provide support for the principle of affirmative action.

1. Introduction

Algorithms are increasingly being used in high-impact domains of decision-making, such as loans, hiring, bail, and university admissions, with wide-ranging societal implications. However, issues have arisen regarding the fairness of these algorithmic decisions. For example, the risk assessment software, COMPAS, used by judicial systems in many states, predicts a score indicating the likelihood of a defendant committing a crime if given bail. ProPublica analyzed recidivism predictions from COMPAS for 10,000 criminal defendants, looked at false positive rates and false negative rates for defendants of different races, and found the tool to be biased against black defendants [3]. Equivant (formerly called Northpointe), the company that developed the COMPAS tool, on the other hand, focused on positive predictive value, which is similar for whites and blacks [12]. That is, by some measures of fairness, the tool was found to be biased against blacks; meanwhile by other measures, it was not. Which definition of fairness should be measured?

[☆] This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

[*] Corresponding authors.
E-mail addresses: nsaxena@usc.edu (N.A. Saxena), karenhuang@g.harvard.edu (K. Huang), defilippis@g.harvard.edu (E. DeFilippis), gradanovic@mpi-sws.org (G. Radanovic), parkes@eecs.harvard.edu (D.C. Parkes), yangliu@ucsc.edu (Y. Liu).

[1] This work was done while the author Goran Radanovic was at Harvard University.

https://doi.org/10.1016/j.artint.2020.103238
0004-3702/© 2020 Elsevier B.V. All rights reserved.

The above scenario is not a rare case. Given the increasing pervasiveness of automated decision-making systems, there's a growing concern among both 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 [17]. It is very unlikely that one definition of fairness will be sufficient. This is supported also by recent impossibility results that show some fairness definitions cannot coexist [23]. We propose that to the extent computer science researchers are interested in public responses to algorithmic definitions of fairness, it would be important to investigate these public views [26,27,25,8,34].

Substantial research has been done within fields such as moral psychology [35,6], law [16,32], and sociology [7] in order to understand people's perceptions of fairness. We contribute to an emerging area of research on how the general public views fairness criteria in algorithmic decision making. Prior research has investigated the influence of different factors on views of algorithmic fairness [30], perceptions of discrimination in targeted online advertising [31], perceptions of features used in algorithmic decision making [20,19], perceptions of justice in algorithmic decision making under different explanation styles [8], and preferences about equality of false positive/negative rates and accuracy in different contexts [33]. In contrast to this work, 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 believe this understanding will open an important conversation between computer science researchers and the public in how precisely to define algorithmic fairness when building AI systems.

We take three definitions of fairness from the computer science literature, and design experiments to study how laypeople perceive these definitions. Overall, within the domain of loan allocations, the results show a public preference for one of these definitions: calibrated fairness. Furthermore, the results show that providing contextual information regarding the recipients of the loan applicants – that is, sensitive information on race or gender – influences respondents' perceptions of fairness. Our results provide some evidence for public support for affirmative action within the domain of loan allocations.

By testing people's perception of different fairness definitions, we hope to spur more work on understanding definitions of fairness that would be appropriate for particular contexts. In line with recent work examining public attitudes of the ethical programming of machines [9,5], we suggest that these public attitudes serve as useful and important input in a conversation between technologists and ethicists. Much of the work in this line of literature has used opinion polling (or more generally crowdsourcing) to examine public attitudes about fairness. As an example, Yaari and Bar-Hillel [35], Pierson [30], Plane et al. [31], Grgić-Hlača et al. [20], Grgic-Hlaca et al. [19], Binns et al. [8] used survey methodologies to examine people's attitudes about fairness and discrimination. These findings can help technologists to develop decision-making algorithms with fairness principles aligned with those of the general public, to make sure that designs are sensitive to the prevailing notions of fairness in society. Crowdsourcing can also be used to understand how preferences vary across geographies and cultures. Though far from perfect, Amazon Mechanical Turk provides one representative population to carry out large-scale online studies in order to conduct research on this domain [29,10].

By showing the general public's attitudes toward three definitions of algorithmic fairness from the computer science literature, our research contributes to an ongoing debate within the computer science community regarding how to define algorithmic fairness. We suggest that it is important to take into account the principle of affirmative action, which seems to matter to the general public, even though the definitions themselves would not have allowed sensitive attributes like gender or race to influence loan allocations. Such discrepancies between the assumptions made by computer scientists and the fairness perceptions of the general public should be thoughtfully considered before deploying tools that may directly affect the public.

2. Definitions of fairness

Broadly, we investigate a concept of fairness known as *distributive justice*, or fairness regarding the outcomes [1,2]. Which characteristics regarding the individual should be relevant and irrelevant to fairness regarding that individual's outcomes? In this study, we focus on investigating two characteristics: task-specific similarity (loan repayment rate) and a sensitive attribute (race or gender). We collect data on attitudes toward the relevancy of these characteristics. In principle, a fair outcome would be determined without any bias regarding an individual's inherent or acquired characteristics, which would be irrelevant to the particular decision-making context [11]. In many contexts, these inherent characteristics (referred to as 'sensitive attributes' or 'protected attributes' in the computer science literature) would be gender, religion, race, skin color, age, and national origin.

We restrict our emphasis to three fairness definitions from the computer science literature. We chose to test these three definitions for several reasons. First, these definitions could be easily operationalized as decisions in the context of loan scenarios that could be understandable by non-experts. Since one challenge in this research is explaining the algorithmic fairness definitions clearly to participants, we decided to focus on the fairness definitions that could be easily conveyed to survey participants, and yet would be rich enough to capture the variety of fairness definitions in the computer science literature. Second, the definitions that we chose focus on the concept of individual fairness, which refers to fairness to all individuals, regardless of their membership in any group. One of our research objectives was to investigate to what extent and in what way a definition related to this concept of individual fairness should be constrained.

In our experiments, we map these definitions (or constrained versions of the definitions) to specific loan allocation choices, and test people's judgments of these choices. We summarize the three fairness definitions as follows:

Treating similar individuals similarly. Dwork et al. [14] formulate fairness as treating similar individuals (with respect to certain attributes) similarly in receiving a favorable decision, where the similarity of any two individuals is determined on the basis of a similarity (distance) metric, specific to the task at hand, which ideally represents a notion of ground truth with regard to the decision context. Given this similarity metric, an algorithm would be fair if its decisions satisfied the Lipschitz condition (a continuity and similarity measure) defined with respect to the metric. In our loan allocation scenario, individuals with similar repayment rates should receive similar amounts of money.

Given a distance or similarity metric, an algorithm would be fair according to Dwork et al. [14] if it satisfied the Lipschitz condition with respect to it. Let V be the set of individuals, A the set of possible outcomes, and let M denote the mapping from the set of individuals to the probability distributions over the possible outcomes. Let d represent the distance or similarity metric $d : V \times V \to \mathbb{R}$. Then $M : V \to \Delta(A)$ will satisfy the (D, d)-Lipschitz property if for every $x, y \in V$,

$$D\big(M(x), M(y)\big) \leq d(x, y)$$

Never favor a worse individual over a better one. In a setting where a single individual is to be selected for a favorable decision, Joseph et al. [22] define fairness 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. This definition promotes meritocracy with respect to the candidate's inherent quality. Joseph et al. [22] apply this definition of fairness to the setting of contextual bandits, a classical sequential decision-making process, by utilizing the expected reward to determine the quality of an action (i.e., an arm as in the bandit setting). Each arm represents a different subpopulation, and each subpopulation may have its own function that maps decision context to expected payoff. In our loan allocation scenario, an individual with a higher repayment rate should obtain at least as much money as her peer.

Joseph et al. [22] apply this definition of fairness to the setting of contextual bandits, by utilizing the expected reward to determine the "quality" of an arm. Each arm represents a different subpopulation, and each subpopulation may have its own function that maps the contexts to the expected payoffs (denoting using f_j for arm j). According to this definition, an algorithm is fair if with high probability, over all the rounds t, and for all pairs of arms j and j', when their expected rewards satisfy that $f_j(x_j^t) \geq f_{j'}(x_{j'}^t)$, it will choose arm j with at least the same probability with which it chooses arm j'.

Let $\pi_{j|h^t}^t$ denote the probability that the algorithm chooses the arm j over the arm j' after observing the contexts x^t, given the history h^t. In the contextual bandits scenario, an algorithm will be δ-fair if for all payoff distributions, rounds $t \in T$, all pairs of arms j, j', and all context sequences $x^1, \ldots, x^t$, with probability at least $1 - \delta$ over the realization of history h^t,

$$\pi_{j|h^t}^t > \pi_{j'|h^t}^t \quad \text{only if} \quad f_j(x_j^t) \geq f_{j'}(x_{j'}^t)$$

In the classic stochastic multiarmed bandits setting, which is a special case of the contextual bandit problem in which the contexts are the same every day, and every arm j has an unknown distribution over $[0, 1]$ with unknown mean μ_j, the fairness constraint requires that with probability $1 - \delta$, for arms j and j' with $\mu_j > \mu_{j'}$, the algorithm never plays arm j' with a probability greater than that of playing j.

Calibrated fairness. The third definition, which we refer to as 'calibrated fairness', is formulated by Liu et al. [28] in the setting of sequential decision-making.[2] Calibrated fairness selects individuals in proportion to their merit. In a multi-armed bandit setting, this means that an arm would be pulled with a probability that its pull would result in the largest reward if all the arms are pulled. When the merit (i.e., underlying true quality) is known, calibrated fairness implies the meritocratic fairness of Joseph et al. [22]. Furthermore, as argued by Liu et al. [28], calibrated fairness implies fairness as defined by Dwork et al. [14] for a suitably chosen similarity metric. In our loan allocation scenario, we interpret calibrated fairness 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.[3]

The calibrated fairness definition as proposed by Liu et al. [28] is a measure of proportionality of the quality of the individuals. It does not refer to calibration within groups. We refer to the definition proposed by Joseph et al. [22] as meritocratic fairness to be consistent with the literature.

3. Overview of present research

In the present research, we investigate the following question: When do people endorse one fairness definition over another in the context of loan allocations?

First, we want to understand how variation in the task-specific similarity of the target individuals may influence people's support for the three definitions of fairness. The three definitions differ in the extent to which comparison between task-specific metrics should matter.

[2] Note that Kleinberg et al. [23], Chouldechova [11] define 'calibration' in a different way, that includes the notion of a sensitive attribute.

[3] This is a slightly different version of the formal definition in [28], 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.

Second, we aim to understand how information about the race or gender of the target individual loan recipients may influence these fairness perceptions. Direct discrimination is the phenomenon of discriminating against an individual simply because of their membership, or perceived membership, in social categories identified by protected (or sensitive) attributes, such as age, disability, religion, gender, and race [15]. All three definitions agree that, conditioned on the relevant task-specific metric, an attribute such as race or gender should not be relevant to decision-making.[4] Information about race or gender may matter, however, since people may consider these to be important factors for distributive justice. For example, in decisions promoting affirmative action, people may believe that considering race (or gender) is important in order to address historical inequities. If that is the case, then definitions of algorithmic fairness may need to take into account such sensitive attributes.

Across three online experiments, we investigate how people perceive algorithmic fairness in the context of loan decisions – a setting where a divisible good must be allocated. We employ a scenario where a loan officer must decide how to allocate a limited amount of loan money to two individuals. In Study 1, we test how the individuals' task-specific similarity (i.e., loan repayment rates) influences perceptions of fairness, in the absence of information about race or gender. In Study 2, along with the individuals' loan repayment rates, we also mention their race and test how that may influence perceptions of fairness. In Study 3, along with the individuals' loan repayment rates, we also mention their gender and test how that may influence perceptions of fairness. Across these studies, we operationalized these fairness definitions, which are formalized for choosing a single individual for a favorable decision (or assigning an indivisible good), to a loan allocation setting where the good is divisible. We limit the scope of our investigation to fairness perceptions within U.S. samples.

4. Study 1 (no sensitive information)

In this study, we aim to investigate how information on an individual task-specific feature (i.e., the candidates' loan repayment rates) 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 rules, described in the following paragraphs, that allow us to formulate qualitative judgments regarding the three fairness definitions.

4.1. Procedure

We recruited 200 participants from Amazon Mechanical Turk (MTurk) on March 18-19, 2018. The majority (82%) of these participants identified as white, 8% identified as black, 6% as Asian or Asian-American, 2% as Hispanic, and the rest identified as multiple races. The average age was 39.43 (SD = 12.47). Most (91%) had attended some college, while almost all other participants had a high school degree or GED. (All demographic information was self-reported.) We collected demographic information from all our participants in order to analyze the representativeness of our results. All participants were U.S. residents, and each were paid $0.20 for participating.

We presented participants with the scenario presented in Fig. 6 in the Appendix. This experiment 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) relevant to the three fairness definitions. That is, we varied the difference in the loan repayment rates of the two individuals across four treatments to help us understand when people perceived two individuals to be similar enough to be treated similarly. In Treatment 1, the difference in loan repayment rates of the two individuals was the smallest: one individual had a loan repayment rate of 55%, and the other had a loan repayment rate of 50%. In Treatment 2, the difference was larger (70% and 40%). The difference was greatest in Treatment 3 (90% and 10%), and Treatment 4 (100% and 20%). Participants were randomly shown one of these four treatments. Each participant was shown only one treatment.

We held all other information about the two candidates constant. We then presented participants with three possible decisions for how to allocate the money between the two individuals. The order of the three decisions was counterbalanced. 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 by Joseph et al. [22], where a worse applicant is never favored over a better one. It would also be allowed under the definition formulated by Dwork et al. [14] 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 [28].

"Equal" Decision. Split the money 50/50 between the candidates, giving $25,000 to each. This decision is allowed in all treatments under the fairness definition formulated by Dwork et al. [14] – that is, treating similar people similarly. Moreover, under their definition, when two individuals are deemed to be similar to each other, then this is the *only* allowable decision (in Treatment 1, for example). This decision is also allowed in all the treatments under the meritocratic definition [22], as

[4] Here, we assume that the treating similar individuals similarly definition [14] does not use race or gender as a relevant dimension for judging individual similarity.

the candidate with the higher loan repayment rate is given at least as much as the other candidate, and, hence, is weakly favored. The decision, however, would not be allowed in any treatment under calibrated fairness [28], 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 [28], 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. [14], 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 [22] for the same reasons as the "Equal" decision. Namely, the candidate with the higher loan repayment rate is weakly favored to the other candidate.

In this research, we test human perceptions regarding the outcomes that different fairness definitions allow, and not the definitions themselves. However, if a certain definition allows multiple decisions, then we would expect these decisions to receive similar support. It would be worthwhile to understand if and when perceptions of the fairness of outcomes may be inconsistent with the allowable decisions for a rule.

If participants most prefer the treating similar people similarly definition, one would expect that they would prefer the "Equal" decision to the other two decisions for a wider range of similarity metrics and treatments. If participants most prefer the meritocratic definition, one would expect no significant difference in support for the three different decisions. If the calibrated fairness definition is the most preferred, one would expect that the "Ratio" decision is perceived as more fair than the other two decisions.

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.

4.2. Results and discussion

First, we tested hypotheses H1A and H1B, which conjecture that participants would consider the "Ratio" decision as the most fair. 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 (see Fig. 1). We found partial support for H1B: Participants rated the "Ratio" decision as more fair than the "All A" decision in Treatments 1 and 2 (see Fig. 1).

Second, we found that participants in Treatment 1 rated the "Equal" decision as more fair than the "All A" definition (see Fig. 1), supporting H2. We observe that when the difference in the loan repayment rates of the individuals was small (5%), participants perceived 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 could be said that participants viewed individuals to be similar enough to be treated similarly only in Treatment 1.

Third, we found that participants rated the "All A" decision as more fair than the "Equal" decision in Treatment 3, but not in Treatment 4 (see Fig. 1).

Evidence from Study 1 suggests that 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 [28] that is based on a notion of merit.

In Treatments 1 and 2, participants rated the "Ratio" decision – the only decision that aligns with calibrated fairness – 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.

5. Study 2 (with 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 provided information on their race and gender. We held the

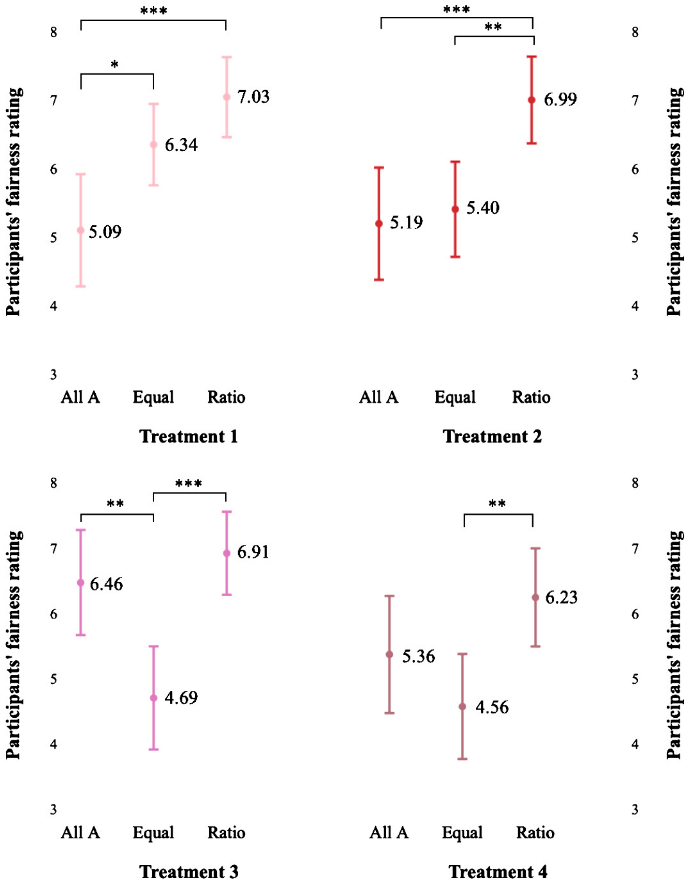


Fig. 1. Comparison of means (with 95% CI) for Study 1. Where * signifies p <0.05, ** p <0.01, and *** p <0.001.

gender of the candidates constant (both were male), and randomized race (black or white). Thus, either the white candidate had the higher loan repayment rate, or the black candidate had the higher loan repayment rate. The question presented to the participants in Study 2 can be found in Fig. 7 in the Appendix.

We recruited a separate sample of 1800 participants from Amazon Mechanical Turk (MTurk) on April 20-21, 2018, none of whom had taken part in Study 1. Most of them identified as white (74%), 9% as black, 7% as Asian or Asian-American, 5% as Hispanic, and the rest with multiple races. The average age was 36.97 (SD = 12.54). Most (89%) had attended some college, while almost all other participants had a high school degree or GED. All participants were U.S. residents, and each was paid $0.20 for participating. (All demographic information was self-reported.)

5.1. Results and discussion

We found that participants viewed the “Ratio” decision as more fair than the “Equal” decision in Treatments 2, 3, and 4, regardless of race, in support of H1A. Furthermore, 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 Figs. 2 and 3.) 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.

Furthermore, participants rated the “Equal” decision as more fair than the “All A” decision in Treatment 1, regardless of race, in support of H2 (see Figs. 2 and 3). Thus when the difference between the loan repayment rates was small (like in Treatment 1), participants preferred the “Equal” decision to the “All A” decision. 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 Figs. 2 and 3).

We see further evidence of the effect of race: 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 black (see Fig. 3). By contrast, when the candidate with the higher

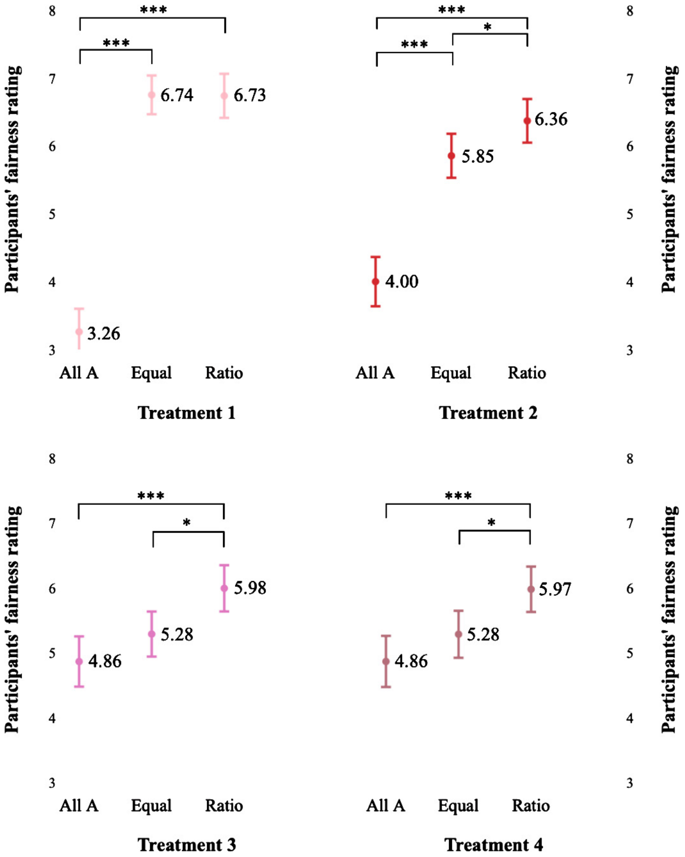


Fig. 2. Comparison of means (with 95% CI) for Study 2 (when the individual with the higher loan repayment rate is white). Where * signifies p <0.05, ** p <0.01, and *** p <0.001.

loan repayment rate was white, participants did not rate the two decisions differently (see Fig. 2). 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 the candidate with the higher payback rate is a member of a group that is historically disadvantaged.

Each definition, from meritocratic to similarity to calibrated fairness is successively stronger in our context, ruling out additional decisions. In this light, it is interesting that the "Ratio" decision is generally most preferred, providing support for the calibrated fairness definition, even though this definition is the strongest of the three in the present context.

6. Study 3 (with sensitive information on gender)

In this study, we investigate if mentioning a different sensitive attribute (gender), 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. However, in this study we held the race of the candidates constant (both were white), and randomized gender (male/female). The question presented to the participants in Study 3 can be found in Fig. 8 in the Appendix.

We recruited a separate sample of 1800 participants from Amazon Mechanical Turk (MTurk) on February 27 - March 2, 2019, none of whom had taken part in Study 1 or Study 2. Most participants (74.4%) identified as white, 9.8% as black, 4.8% as Hispanic, 3.8% as Asian, 2.2% as Asian-American, 0.1% as Native Hawaiian or other Pacific Islander, and the rest with multiple races. The average age was 38.66 (SD = 13.20). Most (89.11%) had attended some college, while almost all other participants had a high school degree or GED. 56.9% identified as female, 43% as male, and 0.1% as non-binary or 'other'. All participants were U.S. residents, and each was paid $0.20 for participating. (All demographic information was self-reported.)

6.1. Results and discussion

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. Furthermore, we found 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

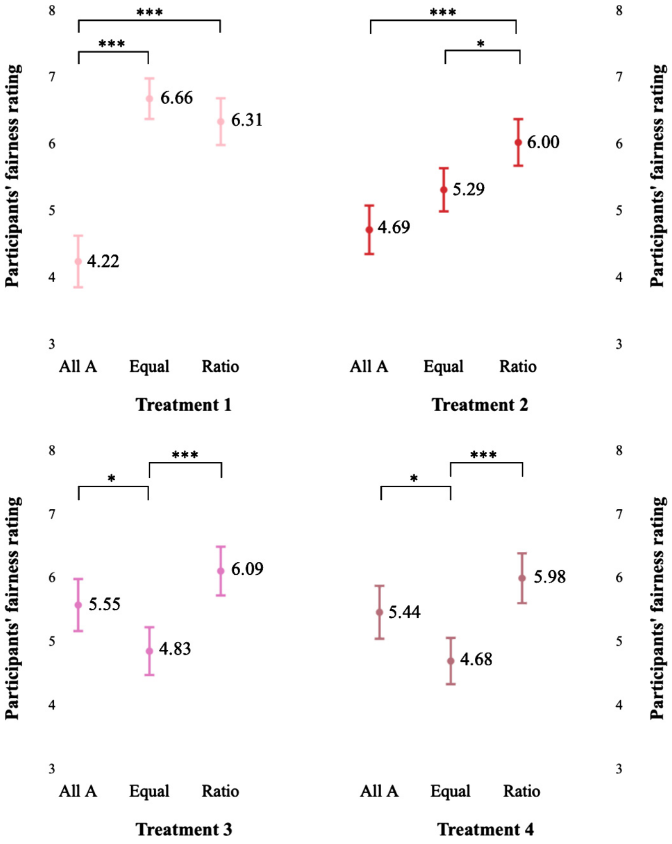


Fig. 3. Comparison of means (with 95% CI) for Study 2 (when the individual with the higher loan repayment rate is black). Where * signifies p <0.05, ** p <0.01, and *** p <0.001.

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 in Treatments 1, 2, and 3, but not in Treatment 4.

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 Figs. 4 and 5). This supports the corresponding results from Study 1 and Study 2. Interestingly, we found that 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 (see Fig. 5).

We find further effects of gender on perceived fairness of the loan allocations. 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. 4). By contrast, when the candidate with the higher loan repayment rate was male, participants did not rate the two decisions differently (see Fig. 5). These results suggest the same boundary condition with regard to H3 that we saw in Study 2: People show more support to giving all the loan money to the candidate with the higher repayment rate rather than dividing the money equally, when the candidate with the higher repayment rate is a member of a historically disadvantaged group.

7. Conclusion

Across three studies, we find that people broadly show a preference for the "Ratio" decision, which is indicative of their support for the calibrated fairness definition [28], as compared to the treating similar people similarly [14] and meritocratic fairness [22] definitions. Furthermore, we observe that sensitive attributes like race and gender do have an effect on people's perceptions of fairness of the loan decisions, suggesting their support for the principle of affirmative action.

We acknowledge several limitations of these findings. First, in our experimental designs, we limited race and gender to binary categories (e.g., male/female or black/white). These studies are meant to show a proof of principle at the psychological level that information on race and gender do affect participants' fairness judgments. In using an experimental design where a variable such as information on race would need to be manipulated, it is common to operationalize constructs in this way. For example, the literature on gender or racial diversity relies on these binary categories of race and gender [4]. That being said, we fully acknowledge the limitations of this binary assumption. Our experimental designs serve as a starting point, and we believe it is important to test further racial categories, or even to vary race without assuming racial categories.

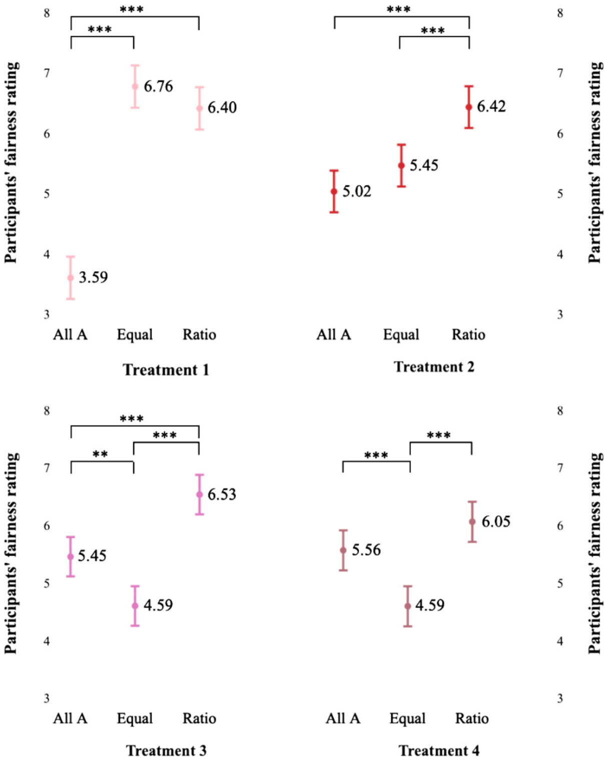


Fig. 4. Comparison of means (with 95% CI) for Study 3 (when the individual with the higher loan repayment rate is female). Where * signifies p <0.05, ** p <0.01, and *** p <0.001.

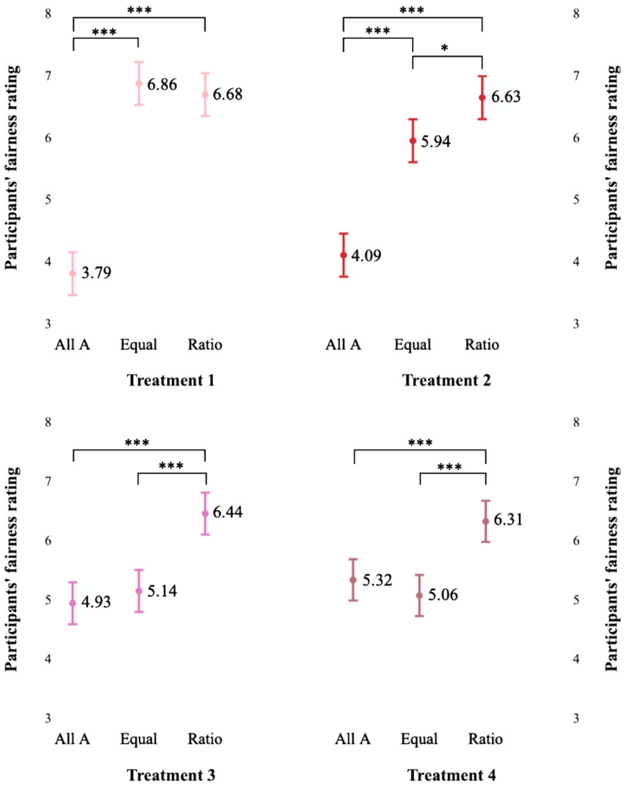


Fig. 5. Comparison of means (with 95% CI) for Study 3 (when the individual with the higher loan repayment rate is male). Where * signifies p <0.05, ** p <0.01, and *** p <0.001.

Second, we note that it is unlikely that one definition of fairness would be sufficient across contexts. Fairness can carry different meanings in different contexts, and our results in the context of allocating divisible goods may not translate to other settings. Even in the scenario of loan decisions, there may be other factors that influence perceived fairness of the decisions (e.g. candidates' incomes, job security, etc.). Different applications may also necessitate different definitions of fairness. We encourage further experimental research testing perceptions of fairness in different settings, in order to investigate the generalizability of these results.

Third, we acknowledge the limitations of Amazon Mechanical Turk as a crowdsourcing platform. Through the use of crowdsourcing, we can elicit information on public attitudes towards different definitions of algorithmic fairness, and how individual characteristics, such as task-specific features (e.g., loan repayment rates) and sensitive attributes (e.g., race or gender) could be relevant for fair decision-making. While we realize that Amazon Mechanical Turk may not be representative of the U.S. population, and serves as a convenience sample, studies have shown that it is more diverse demographically than standard Internet samples tend to be, and much more diverse than typical undergraduate samples tend to be in the U.S. [29,10]. In fact, Paolacci et al. [29] found that MTurk workers tend to be at least as representative of the U.S. population as traditional subject pools. In the future, in order to test for generalizability of our findings, we aim to collect data from further representative sample pools and externally valid field settings.

We emphasize that we neither assume nor demonstrate that our findings on fairness perceptions are desirable. That is, our research is a descriptive project, aimed at capturing people's views of fairness, rather than a normative project. We neither claim that these findings demonstrate moral relativism, nor do we endorse moral relativism. Therefore, these findings do not limit or exclude normative projects on fairness, such as in moral philosophy. Furthermore, we do not claim that the final decision about the most appropriate definition of fairness should solely rely on opinion polls.

That being said, we suggest that these findings, by capturing public attitudes on fairness, can help to continue a dialogue between technologists and ethicists in the design of algorithms that make decisions of consequence for the public. For example, the three fairness definitions examined here agree in the abstract that, conditioned on the task-specific metric, an attribute such as race or gender should not be relevant to decision-making. Yet, we find that under some conditions, people find contextual information on race or gender to be relevant attributes in their fairness perceptions regarding loan decisions. These findings may be relevant for technologists designing these algorithms. Indeed, many of these tools incorporating algorithmic fairness are developed by computer scientists and engineers in the tech industry, who do not represent the general population along metrics of diversity [13,24]. It might be insightful for the developers of such tools to understand how the general public perceives fairness, and to engage in such dialogue. To the extent that multiple stakeholders, such as tech companies, government, and civil society, believe that such a dialogue would be of value, we suggest that opinion polling can be informative regarding which attributes should be perceived as relevant when designing fair algorithms.

This paper opens up several directions for future research. Beyond testing additional definitions, future experiments could also specify whether the decision was made by a human or an algorithm, since psychological theories of mind may influence people's fairness judgments. Second, future work could investigate how people perceive fairness in other contexts, such as university admissions or bail decisions, where the decision at hand concerns a resource that is not easily divisible between two candidates. Third, further research could examine why the availability of additional personal or sensitive information influences perceptions of fairness. Why do people consider factors such as race or gender important for their fairness ratings? And to what extent are people willing to endorse affirmative action in defining algorithmic fairness? Finally, it is important to consider how to incorporate the general public's views into algorithmic decision-making.

These results are only the start of a research program on understanding ordinary people's judgments of definitions of algorithmic fairness. As the literature on moral psychology has shown, people often make inconsistent and unreasoned moral judgments [18]. Indeed, research on moral judgments with regard to the decisions made by autonomous vehicles (the "moral machine") has shown that people approve of utilitarian autonomous vehicles in the abstract, but are unwilling to purchase utilitarian autonomous vehicles for themselves [9]. On the other hand, research in moral psychology shows that people can engage in sophisticated moral reasoning, thinking in an impartial, bias-free way, resulting in moral judgments that favor the greater good [21]. Future research could investigate how moral reasoning interventions could influence people's fairness judgments in the domain of algorithmic fairness.

Declaration of competing interest

There is no conflict of interest.

Appendix A

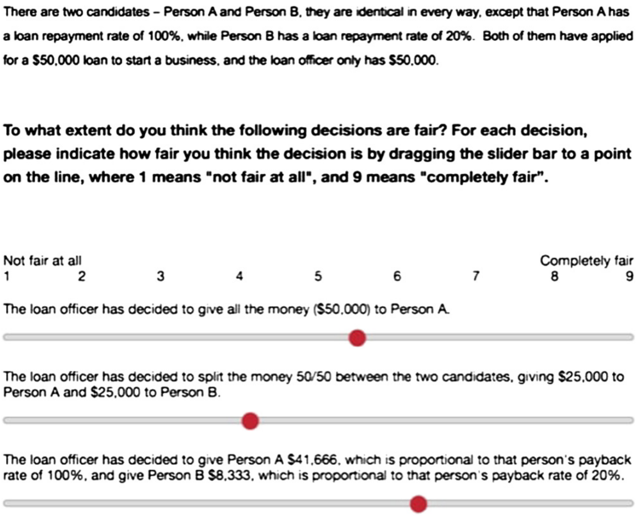
To be eligible to take our surveys, the Amazon Mechanical Turk workers had to be located in the United States of America. We stipulated this restriction via TurkPrime, which is a platform for performing crowdsourced research when using Amazon Mechanical Turk.

Amazon Mechanical Turk workers ('MTurker') could only participate in one of the three studies.

The first section contains the questions the workers were asked in the studies. The second section contains the demographics questions the respondents of all three studies were asked after completing the study. While voluntary, most respondents answered them. The last section contains plots of some demographics of the respondents.

A.1. Questions from the studies

The question asked in Study 1 is presented in Fig. 6. The question asked in Study 2 is presented in Fig. 7. The question asked in Study 3 is presented in Fig. 8.



There are two candidates – Person A and Person B, they are identical in every way, except that Person A has a loan repayment rate of 100%, while Person B has a loan repayment rate of 20%. Both of them have applied for a $50,000 loan to start a business, and the loan officer only has $50,000.

To what extent do you think the following decisions are fair? For each decision, please indicate how fair you think the decision is by dragging the slider bar to a point on the line, where 1 means "not fair at all", and 9 means "completely fair".

Not fair at all — Completely fair

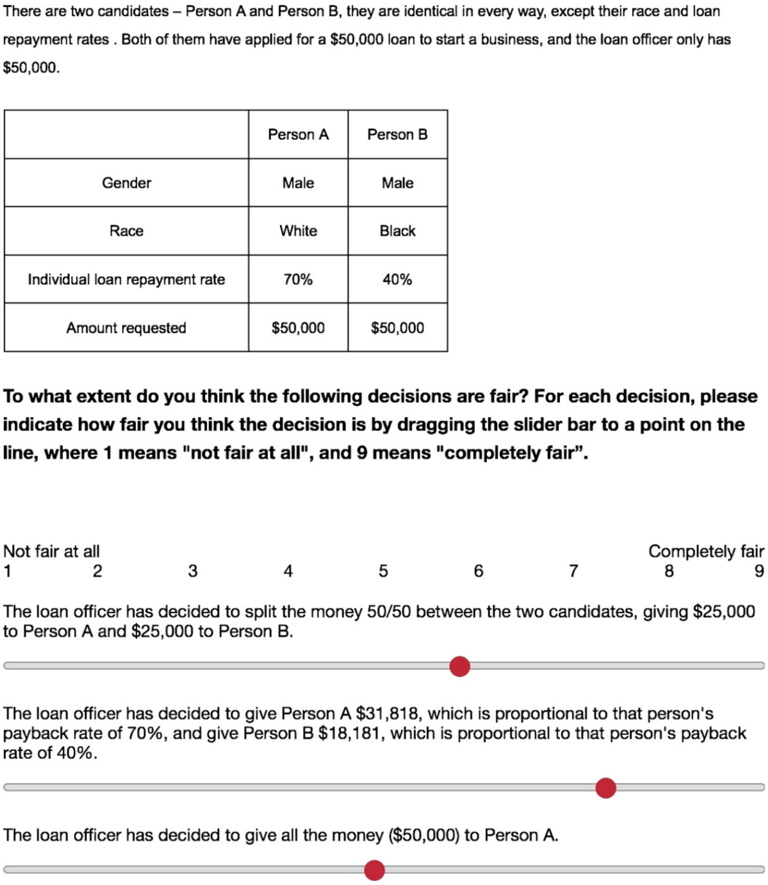
1 2 3 4 5 6 7 8 9

The loan officer has decided to give all the money ($50,000) to Person A.

The loan officer has decided to split the money 50/50 between the two candidates, giving $25,000 to Person A and $25,000 to Person B.

The loan officer has decided to give Person A $41,666, which is proportional to that person's payback rate of 100%, and give Person B $8,333, which is proportional to that person's payback rate of 20%.

Fig. 6. Question presented to the participants in Study 1.



There are two candidates – Person A and Person B, they are identical in every way, except their race and loan repayment rates . Both of them have applied for a $50,000 loan to start a business, and the loan officer only has $50,000.

	Person A	Person B
Gender	Male	Male
Race	White	Black
Individual loan repayment rate	70%	40%
Amount requested	$50,000	$50,000

To what extent do you think the following decisions are fair? For each decision, please indicate how fair you think the decision is by dragging the slider bar to a point on the line, where 1 means "not fair at all", and 9 means "completely fair".

Not fair at all — Completely fair

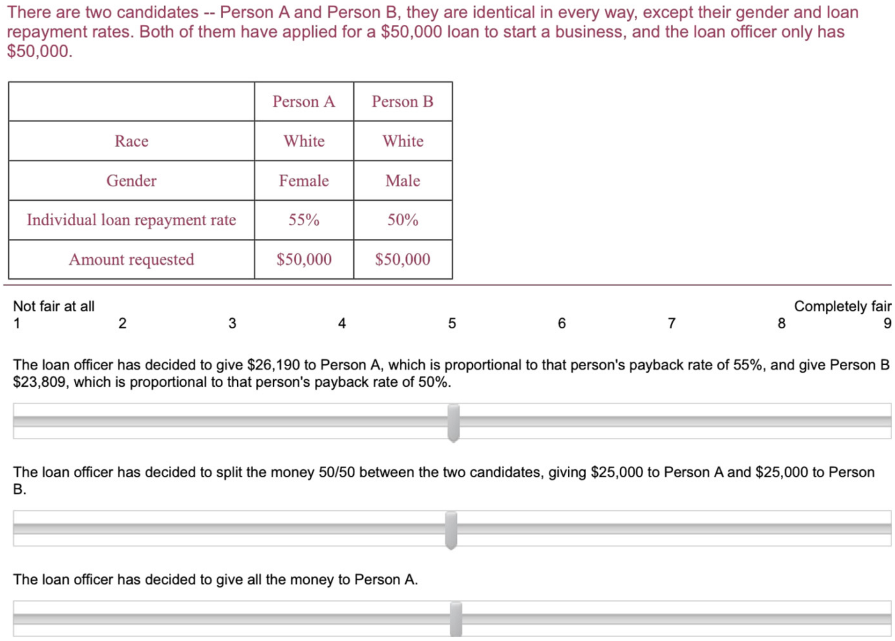
1 2 3 4 5 6 7 8 9

The loan officer has decided to split the money 50/50 between the two candidates, giving $25,000 to Person A and $25,000 to Person B.

The loan officer has decided to give Person A $31,818, which is proportional to that person's payback rate of 70%, and give Person B $18,181, which is proportional to that person's payback rate of 40%.

The loan officer has decided to give all the money ($50,000) to Person A.

Fig. 7. Question presented to the participants in Study 2.



There are two candidates -- Person A and Person B, they are identical in every way, except their gender and loan repayment rates. Both of them have applied for a $50,000 loan to start a business, and the loan officer only has $50,000.

	Person A	Person B
Race	White	White
Gender	Female	Male
Individual loan repayment rate	55%	50%
Amount requested	$50,000	$50,000

Not fair at all ... Completely fair
1 2 3 4 5 6 7 8 9

The loan officer has decided to give $26,190 to Person A, which is proportional to that person's payback rate of 55%, and give Person B $23,809, which is proportional to that person's payback rate of 50%.

The loan officer has decided to split the money 50/50 between the two candidates, giving $25,000 to Person A and $25,000 to Person B.

The loan officer has decided to give all the money to Person A.

Fig. 8. Question presented to the participants in Study 3.

A.2. Demographics questions from the studies

1. What state do you live in?
2. Do you identify as:
   ○ Male
   ○ Female
   ○ Other (please specify): ________
3. What is the highest level of school you have completed or the highest degree you have received?
   ○ Less than high school degree
   ○ High school degree or equivalent
   ○ Some college but no degree
   ○ Associate degree
   ○ Bachelor degree
   ○ Graduate degree
4. Do you identify as:
   □ Spanish, Hispanic, or Latino
   □ White
   □ Black or African-American
   □ American-Indian or Alaskan Native
   □ Asian
   □ Asian-American
   □ Native Hawaiian or other Pacific Islander
   □ Other (please specify): ________
5. In what type of community do you live:
   □ City or urban community
   □ Suburban community
   □ Rural community
   □ Other (please specify): ________
6. What is your age?
7. Which political party do you identify with?
   □ Democratic Party
   □ Republican Party
   □ Green Party
   □ Libertarian Party
   □ Independent
   □ Other (please specify): ________

A.3. Study 1: demographic information of the participants (see Figs. 9 and 10)

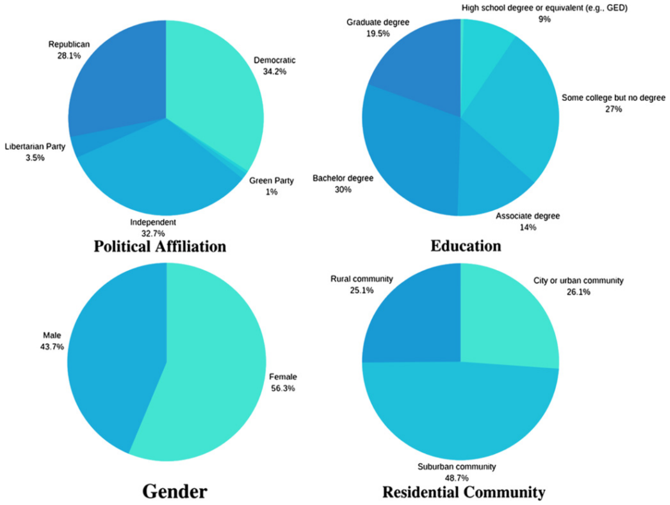


Fig. 9. Political affiliation, education attainment, gender, and residential community of the participants in Study 1.

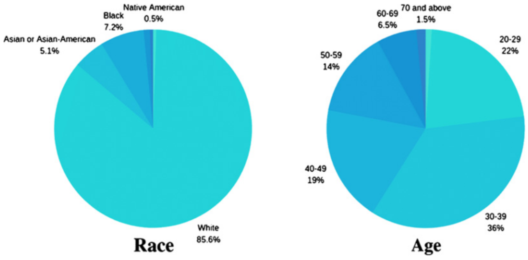


Fig. 10. Race and age of the participants in Study 1.

A.4. Study 2: demographic information of the participants (see Figs. 11 and 12)

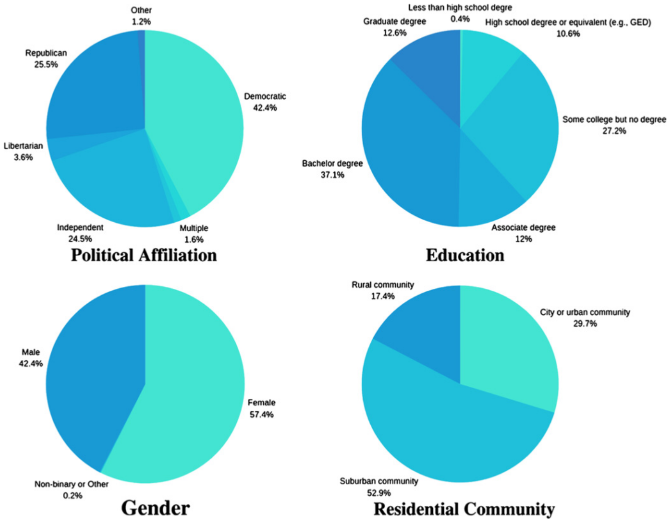


Fig. 11. Political affiliation, education attainment, gender, and residential community of the participants in Study 2.

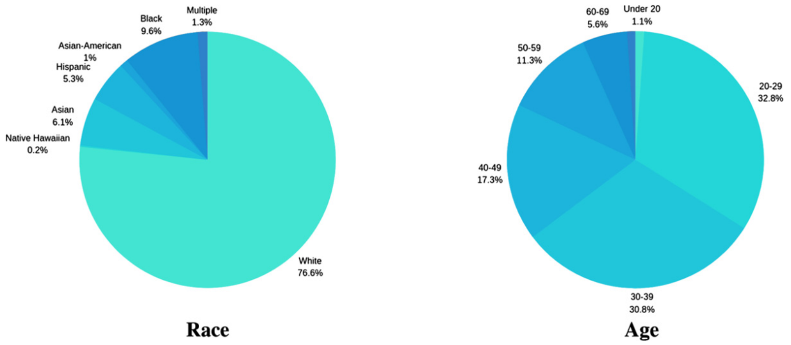


Fig. 12. Race and age of the participants in Study 2.

A.5. Study 3: demographic information of the participants (see Figs. 13 and 14)

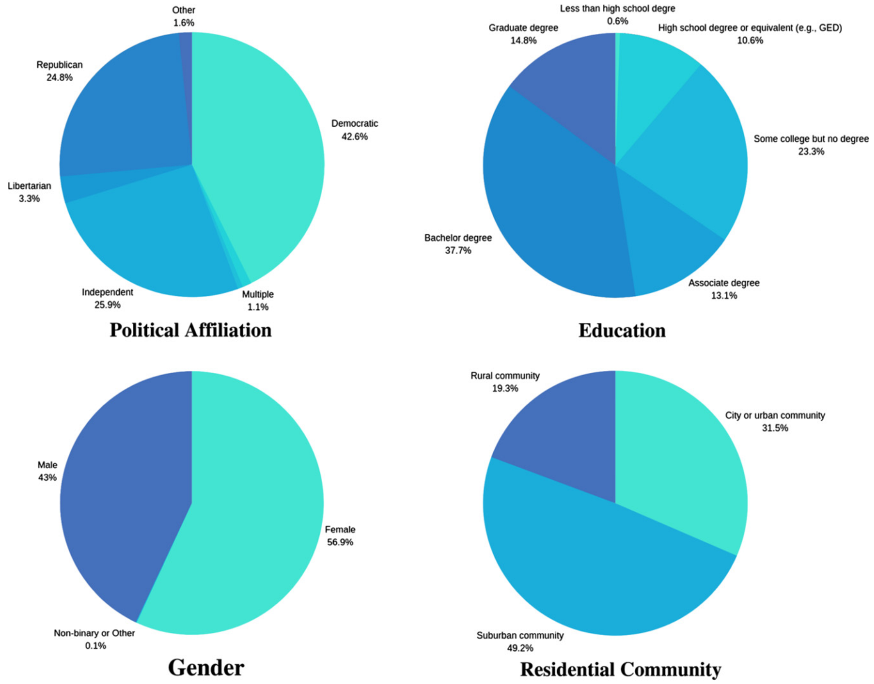


Fig. 13. Political affiliation, education attainment, gender, and residential community of the participants in Study 3.

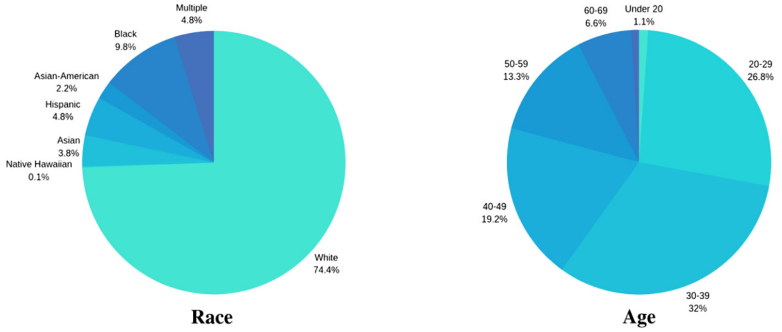


Fig. 14. Race and age of the participants in Study 3.

References

[1] J. Stacy Adams, Towards an understanding of inequity, J. Abnorm. Soc. Psychol. 67 (5) (1963) 422.

[2] J. Stacy Adams, Inequity in social exchange, in: Advances in Experimental Social Psychology, vol. 2, Elsevier, 1965, pp. 267–299.

[3] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias risk assessments in criminal sentencing, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. ProPublica https://www.propublica.org, 2016. (Accessed 27 March 2018).

[4] Anthony Lising Antonio, Mitchell J. Chang, Kenji Hakuta, David A. Kenny, Shana Levin, Jeffrey F. Milem, Effects of racial diversity on complex thinking in college students, Psychol. Sci. 15 (8) (2004) 507–510.

[5] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, Iyad Rahwan, The moral machine experiment, Nature 563 (7729) (2018) 59.

[6] H. Max Bazerman, Sally Blount White, George F. Loewenstein, Perceptions of fairness in interpersonal and individual choice situations, Curr. Dir. Psychol. Sci. 4 (2) (1995) 39–43.

[7] Sebastian Benthall, Bruce D. Haynes, Racial categories in machine learning, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, 2019, pp. 289–298.

[8] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, Nigel Shadbolt, '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, 2018, p. 377.

[9] Jean-François Bonnefon, Azim Shariff, Iyad Rahwan, The social dilemma of autonomous vehicles, Science 352 (6293) (2016) 1573–1576.

[10] Michael Buhrmester, Tracy Kwang, Samuel D. Gosling, Amazon's mechanical turk: a new source of inexpensive, yet high-quality, data?, Perspect. Psychol. Sci. 6 (1) (2011) 3–5.

[11] Alexandra Chouldechova, Fair prediction with disparate impact: a study of bias in recidivism prediction instruments, Big Data 5 (2) (2017) 153–163.

[12] William Dieterich, Christina Mendoza, Tim Brennan, COMPAS risk scales: demonstrating accuracy equity and predictive parity, http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf, 2016. (Accessed 24 October 2018), Northpoint Inc.

[13] E.C. Dillon Jr., J.E. Gilbert, F.L. Jackson, L.J. Charleston, The state of African Americans in computer science-the need to increase representation, Comput. Res. News 21 (8) (2015).

[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, Richard Zemel, Fairness through awareness, in: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ACM, 2012, pp. 214–226.

[15] Evelyn Ellis, Philippa Watson, EU Anti-Discrimination Law, Oxford University Press, 2012.

[16] Sharon Foley, Deborah L. Kidder, Gary N. Powell, The perceived glass ceiling and justice perceptions: an investigation of Hispanic law associates, J. Manag. 28 (4) (2002) 471–496.

[17] Pratik Gajane, Mykola Pechenizkiy, On formalizing fairness in prediction with machine learning, preprint, arXiv:1710.03184, 2017.

[18] Joshua David Greene, Moral Tribes: Emotion, Reason, and the Gap between Us and Them, Penguin, 2014.

[19] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, Adrian Weller, 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, 2018, pp. 903–912.

[20] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P. Gummadi, Adrian Weller, Beyond distributive fairness in algorithmic decision making: feature selection for procedurally fair learning, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[21] Karen Huang, Joshua David Greene, M. Bazerman, Veil-of-ignorance reasoning favors the greater good, Working paper, 2019.

[22] Matthew Joseph, Michael Kearns, Jamie H. Morgenstern, Aaron Roth, Fairness in learning: classic and contextual bandits, in: Advances in Neural Information Processing Systems, 2016, pp. 325–333.

[23] Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan, Inherent trade-offs in the fair determination of risk scores, preprint, arXiv:1609.05807, 2016.

[24] Sheelah Kolhatkar, The tech industry's gender-discrimination problem, New Yorker (2017).

[25] Min Kyung Lee, Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management, Big Data Soc. 5 (1) (2018), 2053951718756684.

[26] Min Kyung Lee, Su Baykal, Algorithmic mediation in group decisions: fairness perceptions of algorithmically mediated vs. discussion-based social division, in: CSCW, 2017, pp. 1035–1048.

[27] Min Kyung Lee, Ji Tae Kim, Leah Lizarondo, 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, ACM, 2017, pp. 3365–3376.

[28] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, David C. Parkes, Calibrated fairness in bandits, preprint, arXiv:1707.01875, 2017.

[29] Gabriele Paolacci, Jesse Chandler, Panagiotis G. Ipeirotis, Running experiments on amazon mechanical turk, Judgm. Decis. Mak. 5 (5) (2010) 411–419.

[30] Emma Pierson, Demographics and discussion influence views on algorithmic fairness, preprint, arXiv:1712.09124, 2017.

[31] Angelisa C. Plane, Elissa M. Redmiles, Michelle L. Mazurek, Michael Carl Tschantz, Exploring user perceptions of discrimination in online targeted advertising, in: USENIX Security, 2017.

[32] David B. Rottman, Adhere to procedural fairness in the justice system, Criminol. Pub. Pol. 6 (2007) 835.

[33] Megha Srivastava, Hoda Heidari, Andreas Krause, Mathematical notions vs. human perception of fairness: a descriptive approach to fairness for machine learning, preprint, arXiv:1902.04783, 2019.

[34] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, Jeffrey Warshaw, A qualitative exploration of perceptions of algorithmic fairness, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, ACM, 2018, p. 656.

[35] Menahem E. Yaari, Maya Bar-Hillel, On dividing justly, Soc. Choice Welf. 1 (1) (1984) 1–24.