Dimensions of Diversity in Human Perceptions of Algorithmic Fairness

Nina Grgić-Hlača
Max Planck Institute for Software Systems, Max Planck Institute for Research on Collective Goods
Germany
nghlaca@mpi-sws.org

Gabriel Lima
KAIST, Institute for Basic Science
Republic of Korea
gabriel.lima@kaist.ac.kr

Adrian Weller
University of Cambridge, Alan Turing Institute
United Kingdom
adrian.weller@eng.cam.ac.uk

Elissa M. Redmiles
Max Planck Institute for Software Systems
Germany
eredmiles@gmail.com

ABSTRACT

A growing number of oversight boards and regulatory bodies seek to monitor and govern algorithms that make decisions about people's lives. Prior work has explored how people believe algorithmic decisions should be made, but there is little understanding of how individual factors like sociodemographics or direct experience with a decision-making scenario may affect their ethical views. We take a step toward filling this gap by exploring how people's perceptions of one aspect of procedural algorithmic fairness (the fairness of using particular features in an algorithmic decision) relate to their (i) demographics (age, education, gender, race, political views) and (ii) personal experiences with the algorithmic decision-making scenario. We find that political views and personal experience with the algorithmic decision context significantly influence perceptions about the fairness of using different features for bail decision-making. Drawing on our results, we discuss the implications for stakeholder engagement and algorithmic oversight including the need to consider multiple dimensions of diversity in composing oversight and regulatory bodies.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**;

KEYWORDS

Algorithmic Fairness, Fairness in Machine Learning, Human-Centered AI, Human Factors in Machine Learning, Human Perceptions of Fairness

ACM Reference Format:
Nina Grgić-Hlača, Gabriel Lima, Adrian Weller, and Elissa M. Redmiles. 2022. Dimensions of Diversity in Human Perceptions of Algorithmic Fairness. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22), October 6–9, 2022, Arlington, VA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3551624.3555306

EAAMO '22, October 6–9, 2022, Arlington, VA, USA

ACM ISBN 978-1-4503-9477-2/22/10.
https://doi.org/10.1145/3551624.3555306

1 INTRODUCTION

Algorithms are increasingly used to assist humans with making decisions, in contexts ranging from granting bail [5] to medical diagnostics [28]. The impact of algorithmic decision-support on human lives has sparked interest in issues of *algorithmic fairness* [1, 5, 8, 32]. Taking a computational approach, the algorithmic fairness community has proposed various notions of fairness and mechanisms to achieve them [10, 22, 26, 45, 48, 60, 73, 96, 107, 108]; yet, it has been shown that some of these notions are mutually incompatible [18, 23, 34, 65] or misaligned with people's perceptions of fairness [93, 97]. As it is typically not possible simply to enforce a computational constraint to ensure fairness, there have been increasing calls for algorithmic oversight that addresses the multifaceted ethical, legal, and policy questions involved in using algorithms to help make decisions, from the viewpoint of a broad range of stakeholders.

To help navigate this complex space of ethical and moral issues in AI, the European Commission has formed the "High-level expert group on AI" [21], while some corporations have sought to implement oversight boards [49, 84, 88]. These boards and working groups are composed of research and industry professionals, legal experts, journalists, and human rights activists, amongst others [12, 21]. Recent research advocates for further democratizing algorithmic oversight by including not only experts but also those affected by the tools, including the broader public, in such discussions [71, 87, 109]. Yet, the selection of individuals to staff such boards and working groups has been controversial, in part because of concerns around biases of those holding positions on the boards [84]. With the growing emphasis on ensuring equity, diversity, and inclusion in research [7], academia [78], industry [38], and beyond, we explore how diversity in experiences and sociodemographics may be relevant to the design, oversight and governance of algorithms utilized in societally consequential domains.

While much prior work emphasizes the importance of diverse opinions in the discourse about algorithmic fairness [31, 59, 104], little research has examined *which* dimensions of diversity are most critical. Social psychology research suggests that demographic factors [35, 39, 100] affect people's moral judgements. In this line, a few studies have investigated how people's demographics may introduce biases into their perceptions of algorithmic fairness [2, 6, 83, 103]. Motivated by the call for inclusion of those affected by

algorithmic decisions to oversight and regulatory bodies, we extend previous research by considering a broader range of demographic factors, as well as individuals' personal experiences, which have also been found to influence their moral views [3, 15, 37, 47, 74, 92].

We study this question in the context of one aspect of procedural fairness—the fairness of using particular features in an algorithm—in a societally consequential scenario: algorithm-assisted bail decision-making. We run a human-subject study ($n = 329$) to evaluate the differences in people's fairness judgements based on (i) demographics (age, education, gender, race, political views) and (ii) personal experience with the algorithmic task being evaluated.

We find that people's political views are associated with their beliefs about fairness. Although respondents across the political spectrum rank algorithmic features—from most to least fair to use—consistently, left-leaning respondents generally consider using an algorithm for bail decisions less fair than their right-leaning counterparts, regardless of the features used. Additionally, we find that people who have had personal experiences that are closely related to the decision-making setting judge the fairness of algorithms using certain features differently than those who did not have such experiences. Namely, the experience of having attended a bail hearing is negatively correlated with the perceived fairness of using information about defendants' juvenile criminal history for making bail decisions.

From this analysis, we provide insight into important dimensions of diversity amongst participants in discussions on algorithmic ethics. These findings offer implications not only for the composition of algorithmic oversight boards and regulatory bodies, but for identifying stakeholders to engage in the design and development of algorithms, composing workshop panels, and evaluating the representativeness of the views presented in conversations about algorithmic fairness more broadly.

2 RELATED WORK

Perceptions of fairness have been extensively studied in social psychology [9, 19, 20, 42, 63, 105]. This field of inquiry has been recently extended to encompass *algorithmic* fairness, as reviewed by Starke et al. [98]. Perceptions of algorithmic fairness have been studied in domains such as targeted advertising [85], lending [93], donation allocation [72], hiring decisions and work evaluation [68, 70], and bail decision-making [43–45, 50, 97]. Most studies have focused on the U.S. population, with few recruiting respondents from other countries, such as Germany [64, 68] and the Netherlands [6].

Many of these studies have found that *people do not reach consensus* in their moral judgments about algorithmic fairness [2, 43, 98]. Prior work studied the possible causes of this variance: properties of the decision context, the algorithmic decision-aid, and people's individual characteristics.

Fairness judgments have been found to vary with the *decision context*. When comparing the perceived fairness of human and algorithmic decisions, Lee [70] found a preference for human decisions in tasks perceived as requiring human skills, such as work evaluation, but no difference in "mechanical tasks", such as work assignment. Nagtegaal [76] found that people perceive human decisions as more fair in high-complexity tasks, while favoring algorithmic ones in low-complexity tasks. On the other hand, Araujo et al. [6] found a preference for algorithmic decisions in high-stakes health and justice decisions.

A different line of research studied how fairness judgments vary with respect to the properties of the *decision-support algorithm*. Shin [94] found a positive correlation between explainability and perceived fairness, and Binns et al. [11] found that this effect depends on the type of explanation provided. Other work focused on the perceived fairness of using specific features for making algorithmic predictions [43–45]. A feature's perceived fairness is positively correlated with its impact on predictive accuracy, perceived relevance, volitionality, and causal relationship with outcomes. On the other hand, features that increase demographic disparity in predictions or are deemed privacy sensitive are perceived as less fair to be used. We build upon this line of work, exploring whether the variance in people's judgments about the fairness of using features can be explained not only by the properties of features, but by people's *individual characteristics*.

Individual Characteristics and Algorithmic Fairness. Araujo et al. [6] studied the association between individual characteristics and the perceived fairness of using algorithms in the health, media, and justice domains. They did not find demographics (age, gender and education) to be significantly associated with fairness judgments. However, the respondents' domain-specific knowledge, beliefs about equality, and online self-efficacy were positively correlated with fairness judgments, while privacy concerns exhibited a negative correlation. Wang et al. [103] studied the fairness perceptions of crowdworkers in the context of an algorithm designed to award Master qualifications on MTurk. They found evidence of egocentric effects, with people perceiving algorithms that assign them negative outcomes as less fair. This effect was stronger for women and less educated participants. However, they did not observe a main effect of demographics (gender, education, age) on fairness judgments but only this interaction effect.

In the present study, we focus on the association between people's individual characteristics and their perceptions of *procedural* fairness, namely the perceived fairness of using different features for making algorithmic decisions. Most closely related to our work, Pierson [83] explored the impact of gender on perceptions of algorithmic fairness, finding women to be less likely than men to favor including gender as a feature in a system designed to recommend courses to students. Grgić-Hlača et al. [45], who considered the same decision context and set of features as we do, descriptively commented on the differences between the extremes on the political spectrum. They found that very liberal respondents perceive the use of features for making bail decisions as less fair than very conservative ones.

Albach and Wright [2] built upon the work of Grgić-Hlača et al. [43], expanding their study of the impact of features' properties on fairness judgments to six distinct decision contexts: bail, child protective services, hospital resources, insurance rates, loans, and unemployment aid. They found that people's judgments were predominantly consistent across the six decision-making domains, with some domain-specific demographic differences. They examined the relationship between three demographic features: gender, race, and educational attainment on these fairness ratings. Across

all six scenarios and features, they found few demographic effects. Of the 4,536 potential relationships they evaluated, only 21 showed a significant correlation with a demographic: race and education. Namely, POC and higher-educated participants rated the fairness of using a few features higher than other respondents. The authors call for further work in this area, given the few relationships they observed and the limited set of factors they analyzed.

In our work, we answer this call. We explore a broader set of individual characteristics, including both a wider set of demographics as well as prior personal experiences, inspired by work in social psychology reviewed below.

Sociodemographics and Moral Judgements. Past research on Moral Foundations Theory [39, 41] has found that sociodemographic features such as gender and political views correlate with people's moral views. Studies indicate that women express more concern about fairness-related moral issues than men [27, 41], and that liberals express more concern about such issues than conservatives [39, 40]. We thus hypothesize similar patterns in our study, looking to see if women and liberals rate certain features as less fair to be used in algorithmic decision-support.

To form hypotheses about political leaning, we can further refer to research on individualist and structuralist beliefs. Compared to liberals, conservatives are more likely to attribute poverty and criminal behavior to individualistic factors—which are under a person's control—than to societal causes, which are beyond one's control [13, 55, 110]. Much prior work has discussed the relationship between the degree of control over outcomes and fairness. Luck egalitarianism argues that people's outcomes should be determined based on their choices, and not on brute luck [4, 66]. Research on the deservingness heuristic has found that people favor allocating social welfare to those they perceive as being unlucky rather than lazy [81]. Finally, Grgić-Hlača et al. [43] found that the perceived volitionality of a feature is positively correlated with its perceived fairness of use in algorithmic decision-support. Hence, as stated above, we hypothesize that conservatism may be positively correlated with the perceived fairness of features.

Additionally, some research found correlations between other sociodemographic factors and perceptions about fairness. African-Americans are more likely to perceive the criminal justice system as unfair [56]. Younger adults are more likely to believe that computer programs can be free from bias [95]. Educational attainment was found to be positively correlated with fairness as considered by Moral Foundations Theory [102]. Hence, we also conduct an exploratory analysis of the association between perceptions of fairness and the respondents' race, age, and education.

Research on egocentric interpretations of fairness [100] suggests that egocentricity may effect people's fairness judgments, especially in individualistic societies [35] such as the U.S. Accordingly, we hypothesize that people's perceptions of the fairness of using specific features, such as age, race or gender, may vary egocentrically based on the respondents' age, race and gender, respectively. Namely, disadvantaged groups (POC, younger individuals, and women[1]) may perceive the use of the corresponding features in algorithmic decision-support as less fair. Alternatively, instead of this ego-justifying and group-justifying behavior, participants may engage in system-justifying behavior [61], studied by System Justification Theory. Disadvantaged individuals and groups are found to sometimes exhibit outgroup favouritism and perpetuate negative stereotypes about themselves [61, 62], while justifying the status quo which puts them in a disadvantaged position. Hence, we alternatively hypothesize that perceptions of fairness may vary in a manner opposite to the egocentric direction, exhibiting outgroup favoritism.

Personal Experiences and Fairness Preferences. Prior work found that people's perceptions of fairness correlate with their past experiences. Namely, negative and traumatic past experiences at both the *individual* and *societal* levels are associated with greater support for fairness interventions.

Alesina and Giuliano [3] found that factors related to a person's past experiences, such as experiencing unemployment and personal traumas, are positively correlated with their support for wealth redistribution. Similarly, Margalit [74] found evidence that economic shocks, such as job loss or a sharp drop in income, tend to increase support for more expansive social policies. Cassar and Klein [15] found that participants who experienced an economic failure in a lab experiment were more likely to favor redistribution, even in the absence of personal monetary stakes. Collectively, these studies show that *individual*-level experiences impact people's perceptions of fairness.

Other studies explored the effects of society-level experiences. Giuliano and Spilimbergo [37] found that an individual's experience of an economic recession while growing up is positively correlated with support for government-led wealth redistribution. In contrast, Roth and Wohlfart [92] showed that people who grew up in times of higher economic inequality are less likely to consider the current real-world income distribution unfair and support wealth redistribution. Gualtieri et al. [47] found that experiencing natural disasters also affects fairness preferences—the intensity of the shakes that people felt during the 2009 l'Aquila earthquake is positively correlated with their support for redistribution.

In our work, we leverage research on the effects of personal experiences and egocentric effects and consider a specific subset of individual characteristics that may exhibit both effects: experiences closely related to the decision-making scenario. We additionally conduct an exploratory study, to explore if respondents who have personal experience with the decision-making task make fairness judgments differently than those who do not.

3 METHODOLOGY

We use a quantitative survey (n = 329) to assess the relationship between respondents' individual factors and their perceptions of procedural fairness in an already well-studied algorithmic support context: bail decision-making. Below, we list our hypotheses, and describe our survey instrument, sampling procedures, analyses, and the limitations of our work. The procedures we describe were approved by our institution's IRB board.

[1]In the context of the COMPAS tool, POC, younger individuals and women can be considered disadvantaged groups, since those individual attributes are significantly correlated with receiving higher risk estimates. Additionally, the COMPAS tool utilized in Broward County, Florida was found to overestimate the criminal recidivism risk of women and POC[69].

3.1 Hypotheses

We leverage prior work reviewed in Section 2 to form the following hypotheses:

Hypothesis 1—Political Leaning: On a scale from very liberal to very conservative, people's political leaning is positively correlated with their fairness ratings.

Hypothesis 2—Gender: Women rate the use of features as less fair than men.

Hypothesis 3a)—Ego- and group-justifying behavior: Disadvantaged groups (women, POC, and younger individuals) rate the use of corresponding features (gender, race, and age respectively) as *less fair* than those who are not members of the disadvantaged group.

Hypothesis 3b)—System-justifying behavior: Disadvantaged groups (women, POC, and younger individuals) rate the use of corresponding features (gender, race, and age respectively) as *more fair* than those who are not members of the disadvantaged group.

We additionally conduct an exploratory study about the relationship between people's perceptions of algorithmic fairness, and a broader set of sociodemographic factors (age, education, and race) and lived experiences.

3.2 Survey Instrument

Survey respondents were presented with an algorithmic bail decision-making scenario inspired by the COMPAS tool, which assists judicial decisions in several U.S. jurisdictions by estimating defendants' risk of criminal recidivism [5]. Using this scenario, we queried respondents' perceptions about the fairness of using eight features from the ProPublica dataset [5], which contains information about more than 7000 criminal defendants who were arrested and subsequently subjected to COMPAS screening in Broward County, Florida in 2013 and 2014. The features about which we asked respondents have received considerable attention from previous work [25, 45, 106] and capture information about the defendants' (1) *number of prior offenses*, (2) *precise description of the current arrest charge*, (3) *degree of the current arrest charge degree*, (4) *number of juvenile felonies*, (5) *number of juvenile misdemeanors*, (6) *age*, (7) *gender*, and (8) *race*.

For each of the eight features, respondents were shown a description of the scenario and asked whether they agreed that it was fair to use the feature for bail decisions using a 7-point Likert scale (1 = "Strongly Disagree," 7 = "Strongly Agree"). The responses to these questions are the study's dependent variables.

Following survey methodology best practices [89], which suggest re-using previously used and already pre-tested survey questions in future research, we draw the phrasing of the scenario and the fairness perception questions from the pre-validated approach of Grgić-Hlača et al. [43]. Figure 2 in the Appendix shows an example vignette for the feature *race*. The eight vignettes were shown in random order to avoid order bias [46, 89]. Respondents then answered two attention-check questions, which we used for quality assurance. The attention-check questions were instructed-response items, in which respondents were instructed to select a specific response option in a multiple-choice question. Similar questions are commonly employed for identifying inattentive respondents in online surveys [75].

Table 1: Demographics of our survey sample, compared to the 2019 U.S. Census [101]. Attributes marked with a † were compared to Pew data [82] on political leaning from 2016.

Demographic Attribute	Sample	Census
<35 years	53.8%	46%
35-54 years	35.6%	26%
55+ years	10.6%	28%
Male	48%	49%
Asian	10.3%	6%
Black	5.8%	12%
Hispanic	7.6%	18%
White	73.6%	61%
Other	2.7%	4%
Liberal / Democrat	50.8%	33%†
Moderate / Independent	24.3%	34%†
Conservative / Republican	24.9%	29%†
Bachelor's or above	60.5%	30%

Table 2: Prior personal experiences of our respondents.

Personal Experiences	Sample
Heard of scenario	6.7%
Legal profession – you	7.6%
Legal profession – friends & relatives	23.1%
Attended bail hearing	14.6%
Served on jury	22.5%

Next, respondents answered a series of questions concerning their personal experiences. Respondents were asked whether they had (1) heard or read anything related to COMPAS before taking the survey; if (2) they or (3) their close friends or relatives held a job or have education in a law or crime-related field; and if they had ever (4) attended a bail hearing or (5) served on a jury. The considered personal experiences vary with respect to their closeness to the task at hand, ranging from close experiences with bail decisions (4), and close experiences with the legal system (2 and 5) to superficial familiarity with the decision context (1) or the legal system (3). Finally, we gathered data about respondents' demographics: age, gender, race, education, and political leaning. The experience questions were shown first in random order, followed by demographic questions shown in random order. Both sets of questions were optional, i.e., respondents had the option to opt out of responding to them. These responses comprise the study's independent variables. The exact phrasing of the aforementioned survey questions is listed in Table 5 in the Appendix.

Finally, we concluded the survey by asking respondents to share their thoughts and feelings about participating in this study,[2] including how interesting they found it, if they would be willing to partake in a similar study in the future, and how difficult they found the questions to understand and respond to.

[2]These questions were inspired by the *enjoyment*, *ease of responding* and *intention to respond to a similar future survey* scales introduced by [24], and used in [52]. For all four questions, we gathered responses using a 5-point Likert scale, from "Strongly agree" to "Strongly disagree'.'

Table 3: Linear mixed model with a random effects term for respondents. Dependent variable: fairness ratings on a 7-point Likert scale with larger values indicating higher perceived fairness. Independent variables (rows): respondents' demographics and experiences. Reference groups for the ordinal variables age and political leaning are "18-24" and "very liberal" respectively. Number of observations = 2632. Standard errors in parentheses. *** p < .001, ** p < .01, * p < .05.

	Coef.	SE
Age	0.0494	(0.0418)
Political leaning	**0.236*****	(0.0454)
Bachelor's or above	0.0297	(0.107)
Male	0.0492	(0.103)
White	-0.0940	(0.116)
Heard of scenario	-0.0312	(0.208)
Legal profession - you	-0.0907	(0.206)
Legal profession - fr & rel	-0.0143	(0.126)
Attended bail hearing	-0.256	(0.148)
Served on jury	0.0432	(0.127)
Constant	**3.668*****	(0.160)

3.3 Sampling

We deployed the survey on the online crowdworking platform Prolific [79]. Prolific is a platform which offers services explicitly targeted at researchers, including a plethora of fine-grained criteria for pre-screening respondents.

Pre-screening Criteria. We used several pre-screening criteria to ensure data quality, per best-practice research guidelines [29]. We targeted respondents who self-reported English fluency and had participated in at least ten previous studies on Prolific, with an approval rate above 95%. Additionally, we only recruited respondents located in the U.S. to ensure that respondents have a basic understanding of the U.S. legal system. Finally, we used additional pre-screening criteria to gather a more representative sample by recruiting more respondents with under-represented demographics and experiences. Namely, we deployed additional studies targeting right-leaning respondents, who are typically under-represented on online crowdworking platforms [53], and men, who are currently under-represented on Prolific [17], as well as people who self-reported to Prolific to have served on a jury, been the victim of a crime, or been to prison. This targeting was not conducted via pre-screening surveys, but instead using Prolific's interface that offers these pre-screening categories, and this sensitive data about participants was not retained nor used in our analysis.

Sample Size Rationale. To estimate the required sample size, we conducted a power analysis using the software G*Power [30]. We accounted for the fact that our sample is likely to be imbalanced in terms of many of the demographics and experiences we consider in our study. While we expected a balanced sample with respect to gender, we anticipated fewer respondents who have served on a jury and even fewer who work in a legal profession. Our goal was to attain a sample that would enable us to identify medium-sized effects (Cohen's $d = 0.5$) for minorities that constitute at least 15% of the sample, and large-sized effects (Cohen's $d = 0.8$) for minorities that constitute at least 5% of the sample. In the context of our correlational study, the effect size captures the magnitude of the difference between the fairness judgments of respondents who self-identify as having different individual characteristics. We focused only on medium- and large-sized effects because small effects may arguably be practically insignificant in our context, regardless of their statistical significance. Based on the two-tailed Wilcoxon-Mann-Whitney test, with standard values $\alpha = 0.05$ and power $(1-\beta)$ 0.8, the minimum required sample size for identifying medium-sized effects for $\geq$ 15% minorities is a total of 260 respondents (39 minority group and 221 majority group). For detecting large effects for $\geq$ 5% minorities, the minimum required sample size is slightly larger: a total of 274 respondents (14 minority group and 260 majority group).

Respondents. We gathered responses from 363 respondents during December 2021. We sampled respondents at different times of the day and on five days of the week to reduce sampling bias that may occur due to the day in the week or the time of day [14]. We removed data from 9 respondents who provided incorrect responses to either of the two attention check questions. Additionally, we discarded all data gathered from 13 respondents who opted not to respond to the demographic or experience-related questions. Finally, we discarded data from 12 respondents who reported their political leaning as "Other," to preserve the ordinal structure of the variable for our analysis. Our final sample consisted of 329 respondents. All respondents were paid 0.8 GBP (approx. $1.08 USD) for completing the study, which took an average of 5.2 minutes. The average hourly rate was hence approximately $12.50.

Table 1 shows the demographics of our sample, compared with the 2019 U.S. Census [101] and 2016 Pew data on political leaning [82]. Compared to the U.S. census, our sample is younger, more educated, more liberal leaning, and consists of more white respondents, as is typically the case for samples recruited on online crowdworking platforms [53, 80, 91]. Table 2 details the respondents' personal experiences related to the decision-making task. Most respondents had not heard of COMPAS nor had they been involved in bail decision-making or juries.

The overwhelming majority of respondents expressed positive sentiments about participating in the study. 93% of respondents found the survey interesting, and 98% stated that they would like to take part in a similar survey in the future. Less than 1% of respondents found the questions difficult to understand, and 4% found them difficult to answer.

3.4 Analysis

We analyzed our data with linear regression models. People's individual characteristics were treated as independent variables, while the 7-point Likert scale fairness ratings were treated as dependent variables. The independent variables are either binary (experiences), ordinal (political leaning: 5-point Likert scale from "Very liberal" to "Very conservative" coded as numerical values from 0 to 4; age: buckets "18-24", "25-34", ..., "85 or older" coded as numerical values from 0 to 7), or transformed to binary (gender: male vs non-male; race: white vs non-white; education: Bachelor's and above vs below

Bachelor's). The specifics of each model are described alongside their results in Section 4.

Limitations. In this paper, we study how people's individual characteristics are associated with their fairness judgments. Individual characteristics, such as demographics and life experiences, are not experimental conditions to which respondents were randomly assigned. Hence, we do not make claims about the causal effects of people's individual characteristics on their fairness judgments but only about the correlation between the two.

Additionally, different individual characteristics are not equally prevalent amongst the respondents. While the fraction of men and women in our sample is balanced, less than 7% of our respondents had heard of the scenario used in the vignettes before participating in this study. This variation in the prevalence of different individual characteristics leads to varying degrees of statistical power to detect effects of interest. In Table 6 in the Appendix, we provide the results of a post-hoc power analysis, which details our sample's statistical power to detect medium and large sized effects. While we have sufficient power to detect large effects for all of the individual characteristics, our study is underpowered (i.e., below the standard $1 - \beta = 0.8$) for detecting medium-sized effects for some of the experience-related characteristics. Hence, one should not interpret the lack of a statistically significant association between an individual characteristic and fairness judgments as evidence that there is no association between the two. It is possible that associations with smaller effect sizes were not identified for the less prevalent, and consequently underpowered personal experiences.

Finally, we studied the dimensions of diversity in fairness perceptions of a sample of U.S.-based respondents for the task of making bail decisions, inspired by the COMPAS tool. We utilize the COMPAS tool as a case study since it is an example of a societally consequential machine learning algorithm that is applied in the real world. We focus on U.S.-based respondents to ensure that they have a basic understanding of the U.S. legal system in which the COMPAS tool is applied. However, prior research on concepts of diversity recognizes that the relevant dimensions of diversity may vary across contexts [31, 99]. Hence, as a promising direction for future research, we encourage the development of a cohesive theory of human reasoning about algorithmic fairness, including the study of additional decision-making scenarios and non-U.S. populations.

4 RESULTS

When asked to determine if it is fair to use the eight ProPublica features for making bail decisions, respondents provided an average response of 4.06 out of 7—close to the midpoint of the Likert scale. While the respondents' fairness ratings averaged across features are neutral, they differ greatly between features, as shown in Figure 1a. In line with the findings of Grgić-Hlača et al. [45], features directly related to the bail decision (current arrest charge and adult criminal history) are considered largely fair to be used. In contrast, distantly related features (juvenile criminal history) are considered less fair, and unrelated sensitive features (age, gender, race) are perceived as unfair for making bail decisions. These descriptive observations are further corroborated by the regression in Table 4, where the constant terms (i.e., the estimated y-axis intercepts) vary between a minimum of 1.078 for race and a maximum of 5.673 for charge description.

4.1 Average Pattern Across Features

We first examine the association between respondents' individual characteristics and perceptions of fairness averaged across all eight features. We employ a mixed-effects linear regression model with fairness ratings as the dependent variable and the respondents' demographics and personal experiences as independent variables. To control for asking each respondent about all eight features (i.e., for repeated measures) we included a random effects term for respondents.

Table 3 shows the results of this analysis. We find that political leaning is significantly correlated with fairness judgments. On average, left-leaning respondents rated using all features as less fair than right-leaning respondents. The regression model estimates that for each step on the 5-point scale from "Very liberal" to "Very conservative", the average fairness rating increases by 0.236 points, suggesting an approximate 1-point difference between the extremes. Figure 1b illustrates mean perceived fairness by political leaning with the corresponding 95% confidence intervals. The figure shows a clear positive correlation up until very conservative respondents, who are rare in our sample and hence have a much larger standard error. No other individual characteristic was found to be significantly correlated with fairness judgments in this analysis.

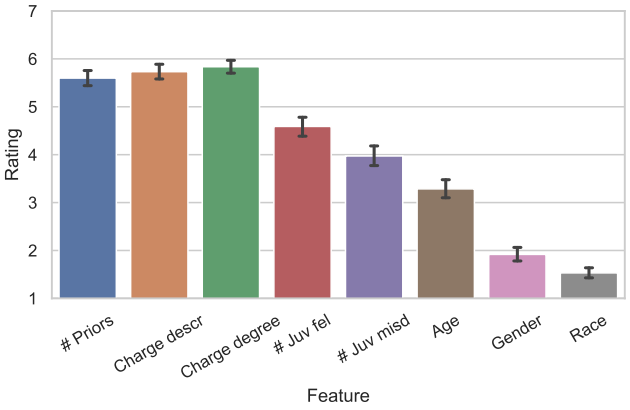
We further explored if the effects of other individual attributes may be subsumed by the effect of political leaning. To investigate this, we trained another model where we removed political leaning from the set of independent variables. We found that in this model the respondents' age exhibits a significant positive association with fairness ratings (coef = 0.086, $p = .045$), while having attended a bail hearing exhibits a borderline significant negative association (coef = -0.3, $p = .051$), as shown in Table 7 in the Appendix.

Summary. When considering the participants' responses averaged across features, we find support for Hypothesis 1 (political leaning), but we find no support for Hypothesis 2 (gender).

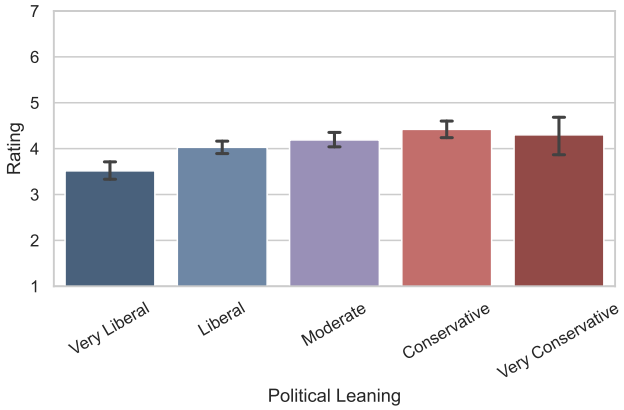
4.2 Feature-Specific Patterns

To analyze the association between respondents' characteristics and their judgments of specific features, we employ a multivariate linear model shown in Table 4. Unlike the model in Section 4.1, which uses a single dependent variable and does not model the variation across features, this model models the eight dependent variables—the fairness ratings of the eight features—simultaneously. Again, we use the respondents' individual characteristics as independent variables.

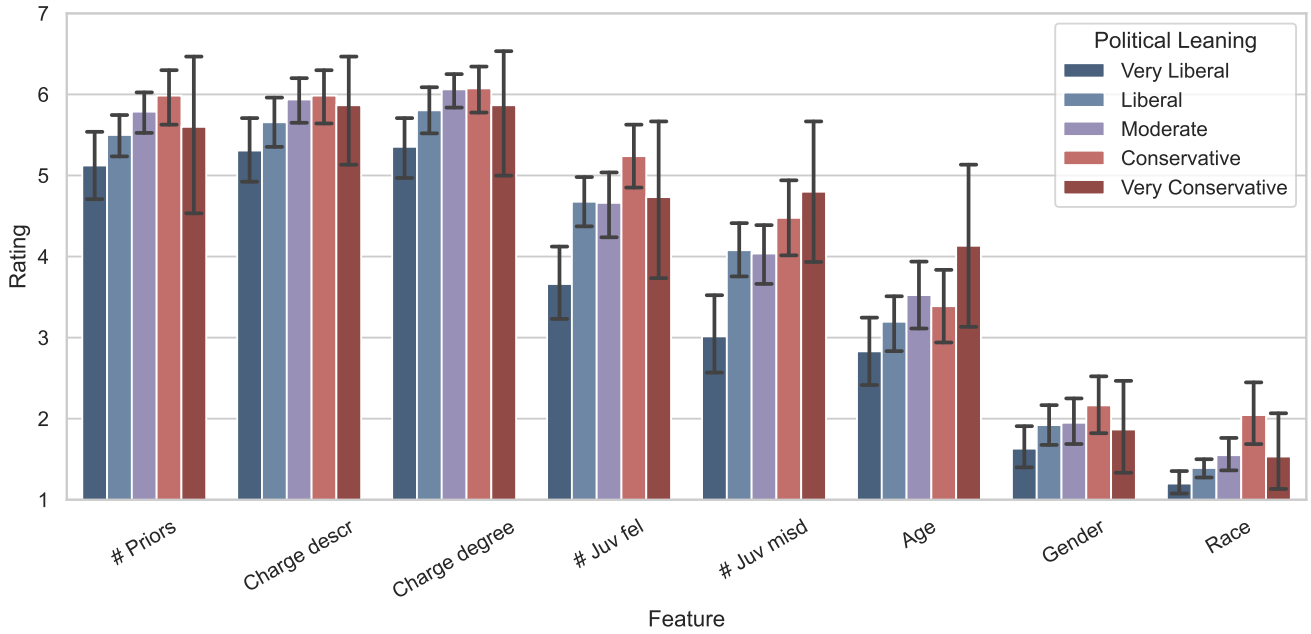
In line with our findings from the previous subsection, political leaning is positively correlated with the perceived fairness of most features. Figure 1c presents this association for each feature separately. We observe that the effect size varies significantly across features. As shown in Table 4, the effect size is the largest and the p-values are the smallest ($p < .001$) for features related to juvenile crimes. For each step on the political spectrum (from "Very liberal" to "Very conservative"), fairness ratings increase by 0.353 and 0.405 points for juvenile felonies and misdemeanors respectively. This corresponds to an estimated difference of approximately 1.5



(a) Mean fairness ratings of the eight ProPublica features.



(b) Mean fairness rating by respondents' political leaning.



(c) Mean fairness ratings of the eight features from the ProPublica dataset by respondents' political leaning.

Figure 1: Mean fairness ratings on a 7-point Likert scale, with 95% CI error bars. Larger values indicate higher perceived fairness.

points between very conservative and very liberal respondents. For the remaining features with a significant association, this effect is smaller, with a difference of approximately 0.8 points between respondents on different ends of the political spectrum. Gender is the only feature not significantly associated with political leaning.

Additionally, we find that having attended a bail hearing is negatively correlated ($p < .01$) with the perceived fairness of using features related to a defendant's juvenile crimes. Respondents who have attended a bail hearing rate the fairness of these features more than two-thirds of a point lower than those who have not. Finally, respondents who identify as men rated using information about a defendant's race by a quarter of a point fairer than others ($p < .05$).

Again, we explored if the respondents' political leaning may be subsuming the effects of their other individual characteristics, by training a model where political leaning is excluded from the set of independent variables. We found that in this model the respondents' age exhibits a significant positive association with the perceived fairness of using information about a defendant's juvenile felonies (coef = 0.195, $p = .015$), as shown in Table 8 in the Appendix.

Table 4: Multivariate linear (structural) model. Dependent variables (columns): fairness ratings of the eight features on a 7-point Likert scale. Larger values indicate higher perceived fairness. Independent variables (rows): respondents' demographics and experiences. Reference groups for the ordinal variables age and political leaning are "18-24" and "very liberal" respectively. Number of observations per dependent variable = 329. Standard errors in parentheses. *** p < .001, ** p < .01, * p < .05.

	# Priors	Charge desc.	Charge deg.	# Juv fel	# Juv misd	Age	Gender	Race
Age	0.0624	0.00239	-0.0402	0.140	0.0337	0.106	0.0797	0.0103
	(0.0650)	(0.0666)	(0.0603)	(0.0791)	(0.0823)	(0.0823)	(0.0581)	(0.0443)
Political leaning	**0.207****	**0.206****	**0.206****	**0.353*****	**0.405*****	**0.216***	0.104	**0.192*****
	(0.0706)	(0.0723)	(0.0655)	(0.0858)	(0.0894)	(0.0894)	(0.0631)	(0.0482)
Bachelor's or above	-0.163	0.0748	-0.0740	0.00458	0.140	0.180	0.127	-0.0514
	(0.167)	(0.171)	(0.155)	(0.203)	(0.211)	(0.211)	(0.149)	(0.114)
Male	-0.00124	-0.298	0.00316	0.180	0.0986	0.0627	0.0821	**0.267***
	(0.160)	(0.164)	(0.148)	(0.194)	(0.203)	(0.202)	(0.143)	(0.109)
White	-0.0596	-0.214	-0.110	-0.412	-0.231	0.0530	0.169	0.0533
	(0.181)	(0.185)	(0.168)	(0.220)	(0.229)	(0.229)	(0.162)	(0.123)
Heard of scenario	0.0939	-0.278	0.254	0.246	0.0289	-0.288	-0.159	-0.148
	(0.322)	(0.330)	(0.299)	(0.392)	(0.409)	(0.408)	(0.288)	(0.220)
Legal profession - you	0.00536	0.126	-0.279	-0.178	-0.00699	-0.0242	-0.166	-0.203
	(0.319)	(0.327)	(0.296)	(0.388)	(0.405)	(0.404)	(0.286)	(0.218)
Legal profession - fr & rel	-0.0499	-0.0363	0.0422	-0.251	-0.0347	0.0217	-0.0649	0.258
	(0.196)	(0.201)	(0.182)	(0.238)	(0.248)	(0.248)	(0.175)	(0.134)
Attended bail hearing	-0.317	0.0278	-0.110	**-0.824****	**-0.700***	0.0506	-0.0485	-0.131
	(0.231)	(0.236)	(0.214)	(0.280)	(0.292)	(0.292)	(0.206)	(0.157)
Served on jury	-0.0288	-0.0115	0.199	0.153	0.150	-0.000934	-0.0251	-0.0906
	(0.198)	(0.203)	(0.183)	(0.241)	(0.251)	(0.250)	(0.177)	(0.135)
Constant	**5.367*****	**5.673*****	**5.665*****	**4.149*****	**3.386*****	**2.597*****	**1.431*****	**1.078*****
	(0.249)	(0.255)	(0.231)	(0.303)	(0.315)	(0.315)	(0.223)	(0.170)

Finally, it is worth noting that while individuals with different political views differ systematically in their absolute fairness assessments, the order in which they rank algorithmic features—from most fair to use to least fair—appears consistent across different political groups. Figure 3 in the Appendix shows the similarity between the rankings by respondents with different political leanings. The rankings are derived from the features' mean fairness ratings by political leaning, and the similarity is quantified using Kendall's Tau (τ). We observe that all pairs of political leanings exhibit a high correlation in their rankings of features, with Kendall's τ values close or equal to 1. That is, across the political spectrum, respondents perceive using information about the defendant's current charge as more fair than using information about the their juvenile criminal history.

Summary. We find support for Hypothesis 1 (political leaning), and weak partial support for Hypothesis 2 (gender). We find no support for Hypotheses 3a) (ego- and group- justifying behavior) and 3b) (system- justifying behavior).

5 DISCUSSION

Here, we discuss our key findings related to the dimensions of diversity explored in this work – political views, demographics, and personal experiences – as well as the implications of these findings.

Political Views. Consistent with prior findings in Moral Foundations Theory [27, 39] and research on diversity in perceptions of algorithmic fairness [45], the respondents' political views were found to be significantly correlated with their fairness judgments for most features. The more conservative an individual is, the more fair they perceive using most features for bail decisions. This trend of conservatives to view information about individuals as fair to use in making decisions about them also aligns with the framework of individualist vs. structuralist beliefs, which has been primarily explored in studies of racism, poverty, and crime in the U.S. [13, 55, 110]. Conservatives tend to believe in "individualist" explanations for outcomes—which emphasize individual responsibility—as compared to liberals, who tend to make structuralist attributions, emphasizing how social structures create outcomes.

It is worth noting that the magnitude and statistical significance of this association varies across features. The effect is the largest for features related to a defendant's juvenile criminal history. This observation may partially be explained by the deservingness heuristic. People's welfare allocation preferences are found to vary depending not only on their political values, but also on the welfare recipient's perceived deservingness. When information about a recipients dersevingness of welfare is available, it is found to outweigh the impact of political values [81]. Some of the features we consider—such as race and gender—are closely related to perceptions of deservingness in welfare allocation settings [58]. Other features—such as information about the current charge degree and adult criminal history—may be closely related to perceptions of deservingness in the task at hand, since they are perceived as the most relevant,

reliable and fair features to be used in the bail decision-making setting [43]. This may explain why the perceived fairness of these features exhibits a smaller association (or, for gender, no association) with the respondents' political leaning.

Demographic Factors. We found no evidence of demographic factors having a consistent significant association with algorithmic fairness judgments. Our finding is in line with the work on perceptions of algorithmic fairness by Araujo et al. [6] and Wang et al. [103], who also found little effect of demographic factors.

The significance of political views and the lack of support for other demographic factors is in line with reports from the Pew Research Center, who find the same patterns in the context of predictors of political attitudes in the U.S [16]. Recent work by Iyengar et al. [57] on affective polarization argues that "[a]s partisan and ideological identities became increasingly aligned, other salient social identities, including race and religion, also converged with partisanships". Hence political leaning may in fact be subsuming other sociodemographic dimensions. Some of our exploratory analyses hint that this may be the case. We found that when we do not control for the respondents' political leaning, their age exhibits some association with fairness ratings. This pattern may be explained in part by the correlation between the respondents' age and political leaning in our sample.

Building upon the social science literature, we also hypothesized that respondents' demographics may relate to the perceived fairness of using those same demographics—age, gender and race—in the decision-making task. We hypothesized that the direction of this effect may be egocentric, in line with research on egocentric interpretations of fairness [100], or the opposite, in line with System Justification Theory [61]. Our results do not offer support for either of these hypotheses. These findings are contrary to those of Pierson [83], who found an ego- and group-justifying association for gender, albeit in a different setting: algorithms for recommending college courses. This may suggest that such associations are scenario-dependent, and future work may seek to explore the relationship between AI fairness beliefs and sociodemographics in other contexts.

Nonetheless, we identified one weakly significant association between demographics and the perceived fairness of using certain features. Namely, we found that men perceive using information about race to be more fair than women do. This finding is in line with prior research on Moral Foundations Theory, which found that women are more concerned about harm and fairness issues than men, even when controlling for their political views [41]. Given that much of modern-day discourse on fairness in policing concerns racial discrimination, especially in the context of algorithms utilized in the law enforcement domain [5, 90], it is possible that race is the most salient fairness issue in the context of our study. Our finding may also be connected to prior research that found that women may exhibit less racial bias than men in certain policy questions [54]. However, these associations were found to be small and inconsistent across studies, perhaps explaining the weakly significant effect we identified.

Personal Experiences. Only one personal experience was found to be significantly associated with fairness judgments: having attended a bail hearing was correlated with a lower likelihood of rating certain features as fair for making bail decisions. Prior research on the effect of past experiences on fairness preferences predominantly focused on the effects of negative and traumatic experiences, such as economic hardships [37, 74] or natural disasters [47]. Having attended a bail hearing is arguably the only of the four experiences that can be classified as negative or traumatic, which may explain why it is the only feature for which we identified an association with fairness judgments.

Additionally, out of the four prior experiences we considered, having attended a bail hearing is the most closely related to the bail decision-making scenario. This suggests that personal experiences may need to be directly tied to the decision at hand in order to lead to an effect. Depending on the respondent's role in the bail hearing [3] this may be an example of an egocentric effect [100] or ego- and group-justifying behavior [61].

Finally, when attending a bail hearing, respondents may have acquired additional information that led them to believe that using information about juvenile criminal history is unfair in such settings, e.g., by learning that juvenile court records have a confidential status in many settings [77]. This raises the question why this pattern was not observed for other legally protected attributes, such as race. A potential explanation could be that the protected status of race is already more broadly applied and widely known than that of juvenile crimes, hence diminishing the learning effect.

Implications. The judgment of which features are fair to use in a given decision-making task is a moral decision, requiring human background knowledge and societal context that may often not be present in the data provided. The appropriate person(s) to make this judgment may vary by setting: policy makers, domain experts, algorithm designers, oversight boards, or even the general population (a descriptive ethics approach [36]).

With increasing calls for diversity in the design, oversight and governance of algorithms, it is important to understand which dimensions of diversity are associated with significant variance across judgments in the context of interest, in order to ensure adequate representation of diverse views. Our results indicate that, at least in the U.S., political leaning may be a critical factor of diversity to consider. This may be timely given observations of increasingly politically homogeneous media outlets [86], corporate boards [33, 51], and academic institutions [67].

Our findings also highlight the importance of considering dimensions of diversity that go beyond demographic and ideological differences. Our work and prior research in human-computer interaction and social psychology all suggest that closely-related personal experiences (e.g., attending a bail hearing) may influence people's moral judgments.

Specifically, in our work, we identified the experience of having attended a bail hearing as one that significantly impacts fairness judgments in the context of machine-assisted bail decisions. However, the relevant prior experiences will likely vary across settings. Further, it is not yet clear when such personal experiences offer benefits—access to otherwise unknown knowledge or context—and when they may lead to a (negative) bias, as in prior work where crowdworkers rated AI that rated them negatively as less fair, even

[3] The question "Have you ever attended a bail hearing?" did not specify the respondent's role in the hearing (e.g., victim, defendant, attorney, ...), in order to minimize the sensitivity and intrusiveness of our experiment.

if the decisions were correct [103]. Thus, further work is needed to better understand in what contexts ensuring diversity with respect to lived experiences is desirable, how closely related personal experiences must be to a given decision context to be relevant, and how the uniqueness of a decision context may influence the relevance of such experiences. We invite future work that moves toward building a cohesive theory that explains the effects of lived experiences on fairness judgments.

Conclusion. The design of ML algorithms aligned with societal moral values is inherently an interdisciplinary endeavor. This requires technical contributions from the ML community, guidelines from policy makers, and research in the social sciences, HCI and related fields that can help gather and understand insights about perceptions of algorithmic fairness. In this work, we identify a set of sociodemographic and experience factors that are associated with people's judgments about the fairness of using various features for making bail decisions. However, relevant dimensions of diversity may differ in other decision-making settings and cultural contexts. We hope this paper will inspire future research about the relevant dimensions of diversity in the governance of AI systems, to help ensure that the diverse perspectives of all stakeholders can be appropriately represented in discussions about societally consequential algorithms.

ACKNOWLEDGMENTS

This research was supported in part by a European Research Council (ERC) Advanced Grant for the project "Foundations for Fair Social Computing", funded under the European Union's Horizon 2020 Framework Programme (grant agreement no. 789373). Adrian Weller acknowledges support from a Turing AI Fellowship under grant EP/V025279/1, The Alan Turing Institute, and the Leverhulme Trust via CFI.

REFERENCES

[1] Amanda Y Agan and Sonja B Starr. 2016. Ban the Box, Criminal Records, and Statistical Discrimination: A Field Experiment. (2016).
[2] Michele Albach and James R Wright. 2021. The Role of Accuracy in Algorithmic Process Fairness Across Multiple Domains. In *Proceedings of the 22nd ACM Conference on Economics and Computation*. 29–49.
[3] Alberto Alesina and Paola Giuliano. 2011. Preferences for redistribution. In *Handbook of social economics*. Vol. 1. Elsevier, 93–131.
[4] Elizabeth S Anderson. 1999. What is the Point of Equality? *Ethics* (1999).
[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[6] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H de Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY* 35, 3 (2020), 611–623.
[7] American Psychological Association. 2021. Equity, Diversity, and Inclusion Framework. (2021). https://www.apa.org/about/apa/equity-diversity-inclusion/framework (Accessed on 08/03/2022).
[8] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* (2016).
[9] Max H Bazerman, Sally Blount White, and George F Loewenstein. 1995. Perceptions of fairness in interpersonal and individual choice situations. *Current Directions in Psychological Science* 4, 2 (1995), 39–43.
[10] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* (2018).
[11] 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*. 1–14.
[12] Oversight Board. 2020. Announcing the First Members of the Oversight Board. https://www.oversightboard.com/news/announcing-the-first-members-of-the-oversight-board/. (Accessed on 02/25/2022).
[13] John S Carroll, William T Perkowitz, Arthur J Lurigio, and Frances M Weaver. 1987. Sentencing goals, causal attributions, ideology, and personality. *Journal of personality and social psychology* 52, 1 (1987), 107.
[14] Logan S Casey, Jesse Chandler, Adam Seth Levine, Andrew Proctor, and Dara Z Strolovitch. 2017. Intertemporal differences among MTurk workers: Time-based sample variations and implications for online data collection. *Sage Open* 7, 2 (2017), 2158244017712774.
[15] Lea Cassar and Arnd Klein. 2017. A matter of perspective: How experience shapes preferences for redistribution. (2017).
[16] Pew Research Center. 2019. In a Politically Polarized Era, Sharp Divides in Both Partisan Coalitions. https://www.pewresearch.org/politics/2019/12/17/in-a-politically-polarized-era-sharp-divides-in-both-partisan-coalitions/. (Accessed on 02/20/2022).
[17] Nick Charalambides. 2021. We recently went viral on TikTok - here's what we learned. https://blog.prolific.co/we-recently-went-viral-on-tiktok-heres-what-we-learned/.
[18] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* (2017).
[19] Jason A Colquitt. 2001. On the dimensionality of organizational justice: A construct validation of a measure. *Journal of applied psychology* 86, 3 (2001), 386.
[20] Jason A Colquitt, Donald E Conlon, Michael J Wesson, Christopher OLH Porter, and K Yee Ng. 2001. Justice at the millennium: a meta-analytic review of 25 years of organizational justice research. *Journal of applied psychology* 86, 3 (2001), 425.
[21] European Commission. 2018. AI HLEG - steering group of the European AI Alliance. https://ec.europa.eu/futurium/en/european-ai-alliance/ai-hleg-steering-group-european-ai-alliance.html. (Accessed on 02/20/2022).
[22] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv preprint arXiv:1808.00023* (2018).
[23] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *KDD*.
[24] Anne-Marie Croteau, Linda Dyer, and Marco Miguel. 2010. Employee reactions to paper and electronic surveys: An experimental comparison. *IEEE Transactions on Professional Communication* 53, 3 (2010), 249–259.
[25] Julia Dressel and Hany Farid. 2018. The Accuracy, Fairness, and Limits of Predicting Recidivism. *Science Advances* (2018).
[26] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 214–226.
[27] Leah Efferson, Andrea Glenn, Rheanna Remmel, and Ravi Iyer. 2017. The influence of gender on the relationship between psychopathy and five moral foundations. *Personality and mental health* 11, 4 (2017), 335–343.
[28] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (2017), 115–118.
[29] Peer Eyal, Rothschild David, Gordon Andrew, Evernden Zak, and Damer Ekaterina. 2021. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* (2021), 1–20.
[30] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.
[31] Sina Fazelpour and Maria De-Arteaga. 2022. Diversity in sociotechnical machine learning systems. *Big Data & Society* 9, 1 (2022), 20539517221082027.
[32] Anthony W. Flores, Christopher T. Lowenkamp, and Kristin Bechtel. 2016. False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks.". (2016).
[33] Vyacheslav Fos, Elisabeth Kempf, and Margarita Tsoutsoura. 2021. The political polarization of US firms. *Available at SSRN 3784969* (2021).
[34] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of Fairness. *arXiv:1609.07236* (2016).
[35] Michele J Gelfand, Marianne Higgins, Lisa H Nishii, Jana L Raver, Alexandria Dominguez, Fumio Murakami, Susumu Yamaguchi, and Midori Toyama. 2002. Culture and egocentric perceptions of fairness in conflict and negotiation. *Journal of Applied Psychology* 87, 5 (2002), 833.
[36] Bernard Gert and Joshua Gert. 2017. The Definition of Morality. In *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.
[37] Paola Giuliano and Antonio Spilimbergo. 2014. Growing up in a Recession. *Review of Economic Studies* 81, 2 (2014), 787–817.
[38] Google. 2022. 2022 Diversity Annual Report. (2022). https://about.google/belonging/diversity-annual-report/2022/ (Accessed on 08/03/2022).

[39] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*. Vol. 47. Elsevier, 55–130.

[40] Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology* 96, 5 (2009), 1029.

[41] Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of personality and social psychology* 101, 2 (2011), 366.

[42] Jerald Greenberg and R Cropanzano. 1993. The social side of fairness: Interpersonal and informational classes of organizational justice. *Justice in the workplace: Approaching fairness in human resource management. Hillsdale, NJ: Lawrence Erlbaum Associates* (1993).

[43] Nina Grgić-Hlača, 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 *WWW*.

[44] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making. In *NIPS Symposium on Machine Learning and the Law*.

[45] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In *AAAI*.

[46] Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. 2011. *Survey methodology*. John Wiley & Sons.

[47] Giovanni Gualtieri, Marcella Nicolini, Fabio Sabatini, and Luca Zamparelli. 2018. Natural disasters and demand for redistribution: lessons from an earthquake. (2018).

[48] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of Opportunity in Supervised Learning. In *NeurIPS*.

[49] Brent Harris. 2019. Establishing Structure and Governance for an Independent Oversight Board. https://newsroom.fb.com/news/2019/09/oversight-board-structure/. (Accessed on 09/25/2019).

[50] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 392–402.

[51] Thao Hoang, Phong TH Ngo, and Le Zhang. 2020. Polarized Corporate Boards. *Available at SSRN 3747607* (2020).

[52] Jason L Huang, Nathan A Bowling, Mengqiao Liu, and Yuhui Li. 2015. Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology* 30, 2 (2015), 299–311.

[53] Connor Huff and Dustin Tingley. 2015. "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics* 2, 3 (2015), 2053168015604648.

[54] Michael Hughes and Steven A Tuch. 2003. Gender differences in whites' racial attitudes: Are women's attitudes really more favorable? *Social psychology quarterly* (2003), 384–401.

[55] Matthew O Hunt. 2004. Race/Ethnicity and Beliefs about Wealth and Poverty. *Social Science Quarterly* (2004).

[56] Jon Hurwitz and Mark Peffley. 2005. Explaining the great racial divide: Perceptions of fairness in the US criminal justice system. *The journal of politics* 67, 3 (2005), 762–783.

[57] Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. The origins and consequences of affective polarization in the United States. *Annual Review of Political Science* 22 (2019), 129–146.

[58] Sebastian Jilke and Lars Tummers. 2018. Which clients are deserving of help? A theoretical model and experimental test. *Journal of Public Administration Research and Theory* 28, 2 (2018), 226–238.

[59] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

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

[61] John T Jost. 2019. A quarter century of system justification theory: Questions, answers, criticisms, and societal applications. *British Journal of Social Psychology* 58, 2 (2019), 263–314.

[62] John T Jost and Mahzarin R Banaji. 1994. The role of stereotyping in system-justification and the production of false consciousness. *British journal of social psychology* 33, 1 (1994), 1–27.

[63] Daniel Kahneman, Jack L Knetsch, and Richard H Thaler. 1986. Fairness and the assumptions of economics. *Journal of business* (1986), S285–S300.

[64] Kimon Kieslich, Birte Keller, and Christopher Starke. 2021. AI-Ethics by Design. Evaluating Public Perception on the Importance of Ethical Design Principles of AI. *arXiv preprint arXiv:2106.00326* (2021).

[65] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *ITCS*.

[66] Carl Knight. 2013. Luck egalitarianism. *Philosophy Compass* 8, 10 (2013), 924–934.

[67] Mitchell Langbert. 2018. Homogenous: The political affiliations of elite liberal arts college faculty. *Academic Questions* 31, 2 (2018), 186–197.

[68] Markus Langer, Cornelius J König, and Maria Papathanasiou. 2019. Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment* 27, 3 (2019), 217–234.

[69] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. (2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[70] 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.

[71] Min Kyung Lee, Nina Grgić-Hlača, Michael Carl Tschantz, Reuben Binns, Adrian Weller, Michelle Carney, and Kori Inkpen. 2020. Human-centered approaches to fair and responsible AI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[72] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Ritesh Noothigattu, Daniel See, Siheon Lee, Christos-Alexandros Psomas, et al. 2018. WeBuildAI: Participatory Framework for Fair and Efficient Algorithmic Governance.

[73] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. 2017. Calibrated Fairness in Bandits. *arXiv preprint arXiv:1707.01875* (2017).

[74] Yotam Margalit. 2019. Political responses to economic shocks. *Annual Review of Political Science* (2019).

[75] Adam W Meade and S Bartholomew Craig. 2012. Identifying careless responses in survey data. *Psychological methods* 17, 3 (2012), 437.

[76] Rosanna Nagtegaal. 2021. The impact of using algorithms for managerial decisions on public employees' procedural justice. *Government Information Quarterly* 38, 1 (2021), 101536.

[77] Kara E Nelson. 1997. The release of juvenile records under Wisconsin's juvenile justice code: A new system of false promises. *Marq. L. Rev.* 81 (1997), 1101.

[78] University of Cambridge. 2022. Equality & Diversity. (2022). https://www.equality.admin.cam.ac.uk/ (Accessed on 08/03/2022).

[79] Stefan Palan and Christian Schitter. 2018. Prolific.ac – A Subject Pool for Online Experiments. *Journal of Behavioral and Experimental Finance* (2018).

[80] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making* (2010).

[81] Michael Bang Petersen, Rune Slothuus, Rune Stubager, and Lise Togeby. 2011. Deservingness versus values in public opinion on welfare: The automaticity of the deservingness heuristic. *European Journal of Political Research* 50, 1 (2011), 24–52.

[82] Pew Research Center. 2016. 2016 Party Identification Detailed Tables. http://www.people-press.org/2016/09/13/2016-party-identification-detailed-tables/

[83] Emma Pierson. 2017. Gender differences in beliefs about algorithmic fairness. (2017). arXiv:1712.09124

[84] Kelsey Piper. 2019. Google's brand-new AI ethics board is already falling apart - Vox. https://www.vox.com/future-perfect/2019/4/3/18292526/google-ai-ethics-board-letter-acquisti-kay-coles-james. (Accessed on 09/25/2019).

[85] Angelisa C Plane, Elissa M Redmiles, Michelle L Mazurek, and Michael Carl Tschantz. 2017. Exploring User Perceptions of Discrimination in Online Targeted Advertising. In {*USENIX*} *Security Symposium*.

[86] Markus Prior. 2013. Media and political polarization. *Annual Review of Political Science* 16 (2013), 101–127.

[87] Iyad Rahwan. 2018. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology* 20, 1 (2018), 5–14.

[88] Lee Rainie and Janna Anderson. 2017. Theme 7: The Need Grows for Algorithmic Literacy, Transparency and Oversight.

[89] Elissa M Redmiles, Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. 2017. *A Summary of Survey Methodology Best Practices for Security and Privacy Researchers*. Technical Report.

[90] Rashida Richardson, Jason M Schultz, and Kate Crawford. 2019. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev. Online* 94 (2019), 15.

[91] Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who Are The Crowdworkers?: Shifting Demographics in Mechanical Turk. In *CHI*.

[92] Christopher Roth and Johannes Wohlfart. 2018. Experienced inequality and preferences for redistribution. *Journal of Public Economics* 167 (2018), 251–262.

[93] Nripsuta Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David Parkes, and Yang Liu. 2019. How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. *AIES* (2019).

[94] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146 (2021), 102551.

[95] Aaron Smith. 2018. Public attitudes toward computer algorithms. https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/. (Accessed on 02/20/2022).
[96] Till Speicher, Hoda Heidari, Nina Grgić-Hlača, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual &Group Unfairness via Inequality Indices. In *KDD*.
[97] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. *arXiv preprint arXiv:1902.04783* (2019).
[98] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2021. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *arXiv preprint arXiv:2103.12016* (2021).
[99] Daniel Steel, Sina Fazelpour, Kinley Gillette, Bianca Crewe, and Michael Burgess. 2018. Multiple diversity concepts and their ethical-epistemic implications. *European journal for philosophy of science* 8, 3 (2018), 761–780.
[100] Leigh Thompson and George Loewenstein. 1992. Egocentric interpretations of fairness and interpersonal conflict. *Organizational Behavior and Human Decision Processes* 51, 2 (1992), 176–197.
[101] U.S. Census Bureau. 2019. American Community Survey 5-Year Estimates.
[102] Florian Van Leeuwen, Bryan L Koenig, Jesse Graham, and Justin H Park. 2014. Moral concerns across the United States: Associations with life-history variables, pathogen prevalence, urbanization, cognitive ability, and social class. *Evolution and Human Behavior* 35, 6 (2014), 464–471.
[103] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 2020. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[104] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating systems. https://ainowinstitute.org/discriminatingsystems.pdf.
[105] Menahem E Yaari and Maya Bar-Hillel. 1984. On dividing justly. *Social choice and welfare* 1, 1 (1984), 1–24.
[106] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*. 1171–1180.
[107] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *WWW*.
[108] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*.
[109] A Zimmerman, Elena Di Rosa, and Hochan Kim. 2020. Technology Can't Fix Algorithmic Injustice. *Boston Review* (2020).
[110] Gail Sahar Zucker and Bernard Weiner. 1993. Conservatism and Perceptions of Poverty: An Attributional Analysis. *Journal of Applied Social Psychology* (1993).